DOCUMENT RESUME

ED 295 972 TM 011 778

AUTHOR Schattgen, Sharon F.; Osterl'nd, Steven J.
TITLE Estimating Norm-Referenced Information from a Criterion-Referenced Test: An Application of the ORT ONLY MODEL.
PUB DATE Apr 88
NOTE 94p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Elementary Secc.idary Education; *Equated Scores; *Estimation (Mathematics); Models; National Norms; *Norm Referenced Tests; Raw Scores; Test Interpretation
IDENTIFIERS *ORT ONLY MODEL

ABSTRACT

An investigation is described of the ORT ONLY MODEL, which is one of four models being used to obtain national norm-referenced and local criterion- or objective-referenced information from a single assessment instrument. Raw scores on a norm-referenced test (NRT) were equated to raw scores on an objective-referenced test (ORT). Resultant equating tables were used to estimate norm-referenced scores for examinees taking only the ORT. Preliminary analyses indicated that corresponding ORT and NRT subject tests were similar in terms of content and statistical properties and that correlation coefficients exceeded the minimum levels for ORT-NRT equating set by Chapter I guidelines. Technical considerations, however, indicate that estimated comparable national percentile ranks should be used cautiously. (SLD)

ED295972

Estimating Norm-referenced Information from a Criterion-referenced Test: An Application of the ORT ONLY MODEL

Sharon F. Schattgen

and

Steven J. Osterlind

Center for Educational Assessment
University of Missouri-Columbia

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education

New Orleans, Louisiana, April 1988

ABSTRACT

Estimating Norm-referenced Information from a Criterion-referenced Test: An Application of the ORT ONLY MODEL

Sharon F. Schattgen & Steven J. Osterlind
University of Missouri-Columbia

This paper describes an initial investigation of the ORT ONLY MODEL, one of four models currently being used to obtain both national norm-referenced and local criterion- or objective-referenced information from a single assessment instrument. Using data from a representative sample of examinees, raw scores on a norm-referenced test (NRT) were equated to raw scores on a state-developed objective-referenced test (ORT) using equipercentile procedures. The resultant equating tables were used to estimate norm-referenced scores, estimated comparable national percentile ranks, for examinees taking only the ORT. The estimated comparable national percentile ranks were reported at the individual student level primarily for Chapter I purposes. While preliminary analyses indicated that corresponding ORT and NRT subject tests were similar in terms of content and statistical properties and that their correlation coefficients exceeded the minimum levels suggested for ORT-NRT equating by Chapter I guidelines, technical considerations suggested that the estimated comparable national percentile ranks be used cautiously pending further research. The practical utility of reporting estimated comparable national percentile ranks at the individual student level is discussed. Strengths of this particular application of the ORT ONLY MODEL and recommendations for future research in this area are delineated.

Estimating Norm-referenced Information from a Criterion-referenced Test: An Application of the ORT ONLY MODEL

Sharon F. Schattgen and Steven J. Osterlind
University of Missouri-Columbia

During recent years, many large scale assessment programs have attempted to obtain both national norm-referenced as well as local criterion- or objective-referenced information from a single assessment instrument. This is done to minimize the cost of the testing program as well as to reduce the time required for testing.

Keene and Holmes (1987) described four models--the NRT ONLY, the NRT-BASED, the ORT-BASED, and the ORT ONLY--currently being used for this type of dual purpose testing and offered several recommendations for further research into the question of "whether the information obtained from the four models is psychometrically valid and practically useful" (p. 26). They specifically called for investigations of the utility of using the models to report norm-referenced interpretations at the individual student level.

This paper describes the initial stage of such an investigation of the ORT ONLY MODEL. The procedures used in Missouri to obtain and report norm-referenced data from the state objective-referenced achievement battery and the methods of presenting these data are delineated. In addition, the practical utility of the norm-referenced information at the individual student level and at the aggregate level is discussed. This paper does not present a detailed technical analysis of the norm-referenced data, but general findings are discussed in order to address implications of using the ORT ONLY MODEL.

Models for Obtaining Norm- and Objective-referenced Information

All four models identified by Keene and Holmes (1987) include the administration of either a nationally normed and standardized achievement test (NRT) developed by a commercial publisher, an objective-referenced achievement test (ORT) developed locally by a state or a district (or their contractor), or some combination of both. "The models differ with respect to the amount of customization employed and the design used to produce the norm-referenced information" (p. 7). A brief description of the four models follows.

The first model consists of administering only a norm-referenced test at one or more grade levels locally and is called the NRT ONLY MODEL. The

content of the NRT matches local objectives; performance is reported for both the national curriculum and the objectives common to the NRT and the local curriculum. The NRT ONLY MODEL is cost-effective and efficient in terms of collecting information. However, the objective-referenced information obtained by using this model is not adequate for anything but ancillary information. Thus it is most appropriately used when norm-referenced rather than local objective-referenced information is emphasized.

The second, the NRT-BASED MODEL, consists of the administration of a nationally normed achievement test along with supplemental items measuring local objectives. The supplemental items may augment those included in the NRT or they may replace them. The NRT-BASED MODEL, like the NRT ONLY MODEL, is cost-effective, and yields valid normative and objective-referenced information. It is most often used to enhance local objectives when the match is minimal between the NRT and the objectives. Extra testing time is required, however, if the supplemental items augment rather than replace those in the NRT.

The ORT-BASED MODEL, the third design, features an objective-referenced test used in combination with some portion of a norm-referenced test. Two designs can be used to collect the norm-referenced information: (a) a representative subset of the NRT items is given to all students, or (b) different sections of the NRT are given to different groups of students through matrix sampling. The ORT-BASED MODEL is efficient in terms of collecting norm-referenced information. In addition, the objectives making up the local curriculum are those on which local instructional programs are evaluated. It is expensive, however, in that it involves the development of a completely customized testing program. Another drawback is that the norm-referenced information is less accurate than that offered by the two models previously described. Thus such data are usually reported at only the aggregate level.

The ORT ONLY MODEL is the fourth design for generating national norm-referenced and local objective-referenced information and the one on which this paper focuses. This model utilizes a concurrent administration of an NRT and an ORT to a representative sample of the student population. The scores from the two tests are equated using any one of a number of equating methods. The norm-referenced scores are then estimated for the student population using their objective-referenced scores. Once the equating has been done, only the ORT needs to be administered during subsequent assessments. This makes the model efficient, but it also raises a question about the long-term accuracy of the norm-referenced information. Thus the normative scores are usually reported at the aggregate level. Keene and Holmes (1987) stressed the critical importance of the content match of the two tests and noted the problem of equating two tests of unequal difficulty. They concluded that "any norm-referenced scores computed with the ORT ONLY MODEL must be used with extreme caution" (p. 22).

Background: Technical and Practical Considerations

Previous Applications of Chapter I A2 Model and ORT ONLY MODEL

The interest in obtaining norm-referenced information from a criterion- or objective-referenced test can be traced to the 1970s, when educators sought assessment systems that would yield data required by federally-funded educational programs, particularly Title I (now Chapter I), and data that could be used locally to evaluate curriculum and instruction. Title I officials devised the A2 model, which uses norms derived from equating an ORT to an NRT, to meet both needs (Echternacht, 1980). The ORT ONLY MODEL is similar to the A2 model except that scores are reported at the individual student level in the A2 model but at only the aggregate level in the ORT ONLY MODEL.

There is not a great body of research on ORT-NRT equating. Of the few studies investigating this topic, most were conducted in the context of Chapter I applications of the A2 model. This research has been limited in scope and has not yielded definite findings about whether or how best to obtain NRT data from an ORT.

Roudabush (1975) was one of the earliest to investigate the possibility of obtaining NRT data from an ORT. He used regression procedures to predict norm-referenced scores from a prescriptive reading inventory. Roudabush found over-prediction at the low end and under-prediction at the high end of the score distribution and concluded that there was considerable variation in the accuracy of predicting individual scores.

In a critique of Title I evaluation models, Linn (1978) stated that the NRT and the ORT used in the A2 design should be "highly correlated" and that the minimum correlation of .60 specified by Title I guidelines was "much too lenient" (p. 12). Linn noted that "under equating conditions much better than can generally be expected for Model A2 applications, systematic errors may be introduced simply due to the equating" (p. 14). A simulated data study led him to conclude that it was inappropriate to use data from one test to establish the expected performance level for another test in the context of Title I evaluations.

Fishbein (1978) and Bunch (1982) also pointed out the necessity of a high correlation between the NRT and the ORT if the two were equated for Title I uses. Fishbein warned of the potential sources of equating error represented by floor and ceiling effects, while Bunch suggested that the Title I A2 model yielded "estimates of estimates" (p. 18).

Despite these concerns, the focus shifted from whether norm-referenced information could be obtained from an ORT to how best to obtain it. Storlie (1979) tried to compare four equating methods--linear, normalized, equipercentile and abbreviated equipercentile--under the Model A2 design but found that the correlations between the NRT and the ORT were too low (all were less than .60) to justify equating. In a follow-up study with simulated data, Crane, Prapuolenis, Rice, and Perlman (1981) compared the same four methods and found that linear equating was generally the preferred method. They also found that, in general, as the correlation between the two tests increased, equating error decreased. Crane et al. concluded that, depending on sample size and equating method, the A2 model could be useful if ORT distributions exhibited a relatively normal distribution.

Several researchers have investigated the use of item response theory equating procedures for obtaining norm-referenced information from an ORT. Both Bauer (1979) and Holmes (1980) used Rasch equating to successfully link criterion-referenced test items to a norm-referenced achievement battery. Bauer's study utilized a state assessment instrument, while Holmes' work focused on a Rasch calibrated item bank. In preparing to implement the ORT ONLY MODEL, the Texas Education Agency (TEA) (1986) compared Rasch and equipercentile procedures in order to determine which should be used to equate the statewide test of minimum skills to a nationally normed achievement battery. TEA found no difference in the expected norm-referenced scores produced from the two methods and concluded that the two procedures yielded "virtually identical results" (p. 4). Even though there was no clear statistical basis for choosing one equating method over another, TEA decided to use the Rasch procedure.

These findings taken together suggest that a high correlation between the NRT and the ORT is necessary for a successful equating, but that equating error is difficult to avoid. The method of equating is likely to influence the accuracy of the equated scores, but there is no conclusive evidence from these studies suggesting the use of one method over another.

Limitations of Equating NRT and ORT Scores

The limitations of ORT-NRT equating were addressed by Angoff (1984) in a discussion of equating tests of unequal reliabilities: "When two tests are not interchangeable--for example when their reliabilities are unequal--their scores cannot be 'equated' in any meaningful way" (p. 101). Angoff noted that scores from two nonparallel tests can, however, be made "comparable with respect to a particular group of examinees if their distributions of scores are identical" (p. 128). Comparability will hold with respect to other groups, but only if those "groups are drawn from the same population as the group on which comparability was originally established" (p. 128).

Angoff suggested that if the methods of equating parallel test forms are applied to the problem of obtaining comparable scores (e.g., NRT data from an ORT), two questions should be asked:

> (1) How similar are the tests for which comparable scores are to be developed?
>
> (2) How appropriate is the group on which the table of comparable scores is based when one considers the person or the group for whom the table is to be used? (p. 139)

According to Angoff, after these questions are answered, the use of the comparable scores and the nature of the decisions that they would be used to make should be considered. His comments imply that under certain circumstances ORT-NRT equating is defensible.

Missouri Application of the ORT ONLY MODEL

Missouri Mastery and Achievement Tests

In 1985, the Missouri General Assembly passed the Excellence in Education Act, which requires all local school districts in the state to test students periodically with criterion-referenced tests over specific objectives in language arts, reading, English, mathematics, science, social studies, and civics. This law also requires that a representative sample of students be tested over these objectives each year, with the results being reported to the legislature. Objectives were written for grades two through ten and are called the "Key Skills."

The *Missouri Mastery and Achievement Tests* (MMAT) (Osterlind, 1987) were created especially to measure students' acquisition of the Key Skills. The MMAT battery consists of objective-referenced tests for grades two through ten in four areas: language arts/reading/English, mathematics, science, and social studies/civics. There are at least two equivalent forms for each test. Every effort was made during the development of the MMAT to adhere to the *Standards for Educational and Psychological Testing* (American Psychological Association, [APA], 1985) in order to ensure that it would yield valid measures of student achievement. Appendix A presents a technical summary of the MMAT.

The first administration of the MMAT, using Form A, occurred in the spring of 1987 to students in grades three, six, eight, and ten. The first administration of the entire battery of tests (Forms B and D) for grades two through ten will occur in the spring of 1988.

Decision to Obtain NRT Data from MMAT

The Missouri Department of Elementary and Secondary Education decided that the MMAT should yield norm-referenced information as well as objective-referenced information, so that local districts could save time and money by administering only one achievement battery. The primary impetus for the Department's decision was districts' need to obtain national norm-referenced information in reading, language arts, and mathematics for all students to determine eligibility for Chapter I services and on Chapter I participants to evaluate the program. The Department also hoped that MMAT norm-referenced information, especially when aggregated, might satisfy other district needs for NRT data.

If norm-referenced information were not available from the MMAT, districts would be forced to administer both an NRT and the MMAT at a minimum of four grades. Districts are required to test students over the Key Skills at a minimum of four grades each academic year--two nonconsecutive levels within the grade span two through six and two nonconsecutive levels within the grade span seven through twelve (Missouri Department of Elementary and Secondary Education, 1986). (Many districts plan to go beyond this requirement and will administer the MMAT to grades two through ten each year.) The limitations and concerns cited previously (e.g., Angoff, 1984; Linn, 1978) were carefully considered before this decision was reached, but ultimately practical considerations outweighed concerns about possible technical problems. It was hoped, however, that by watching out for potential pitfalls and by heeding Angoff's (1984) comments regarding score derivation and use, norm-referenced information could successfully be obtained from the MMAT.

Selection of ORT ONLY MODEL and NRT

The Missouri program emphasizes an objective-referenced assessment of the Key Skills and includes but is not limited to collection of data for a sample of representative students. These features made the ORT ONLY MODEL the appropriate mechanism for obtaining NRT information. Form G of both the *Iowa Tests of Basic Skills* (ITBS) (Hieronymus and Hoover, 1986a) and the *Tests of Achievement and Proficiency* (TAP) (Scannell, 1986a) were chosen for equating to the MMAT at grades two through eight and at grades nine and ten, respectively. The technical characteristics of the ITBS and the TAP are given in the *Preliminary Technical Summary* (Riverside Publishing Co., 1986a).

These two vertically linked NRTs were chosen primarily because they measure content similar to the Key Skills, a critical factor in the success of an ORT-NRT equating (Keene and Holmes, 1987). For information on the

content match between the MMAT and the ITBS/TAP, refer to ITBS/TAP Correlated to Key Skills for Missouri Schools (Riverside Publishing Co., 1986b).

Selection of Equating Method

Equipercentile equating was selected for the Missouri application of the ORT ONLY MODEL because it appeared to be the most appropriate method in light of practical and technical considerations. ITBS/TAP national percentile ranks are derived from raw score tables, so the method selected had to utilize raw scores. Thus, it was not possible to equate using the item response theory two-parameter logistic model (which might have been worthy of consideration had it been supported in the literature) that was used to derive MMAT subject scaled scores. (See Appendix A for a description of MMAT scores and scaling procedures.) The problem of access to the ITBS/TAP norms could have been overcome by Rasch equating, used successfully by Holmes (1980) and the Texas Education Agency (1986), but then the item response theory model used for equating would be different from that used for deriving scaled scores.

Linear equating, which utilizes raw scores, was recommended by Crane, Prapuolenis, Rice, & Perlman (1981). This method, however, assumes that the only differences between the distributions of the two tests being equated are the mean and standard deviation (Crocker and Algina, 1986). Equipercentile equating, which also utilizes raw scores, does not make such an assumption. It is "the only way to ensure equivalent scores when the distribution shapes are different is to equate by curvilinear (equipercentile) methods" (Angoff, 1984, p. 88). Skaggs and Lissitz (1986), in a study of four equating methods, found that equipercentile equating was preferable if the psychometric properties of the two tests being equated were different. The results of a 1986 pilot study, in which the ITBS and field test versions of the MMAT were concurrently administered, indicated that the psychometric properties of the two batteries were in fact different. Thus, equipercentile equating seemed to be the method that would best fit the data.

Exceptions to ORT ONLY MODEL

If MMAT results are to be used for Chapter I purposes, norm-referenced information at the individual student level is needed. In order to improve the accuracy of individual student NRT scores, the NRT and the ORT will be equated each year rather than only once. Therefore, the Missouri procedure features two exceptions to the ORT ONLY MODEL: (a) NRT score reporting at the individual student level, and (b) annual recalibration of the NRT scores.

Method

Subjects

Approximately 240,000 students in grades 3, 6, 8, and 10 participated in the first administration of the MMAT during the spring of 1987. At each of these four grade levels, ten percent of the total number of students tested (around 6,000 per grade) were selected for inclusion in the representative state sample using a stratified random cluster sampling technique. The scores achieved by students in the sample were used to report MMAT results to the legislature and in the equating procedure.

Procedure

A single-group rather than an equivalent-groups design was used for the equating in an attempt to minimize equating error. The students making up the state sample for grades 3, 6, and 8 were randomly assigned to one of five groups. One subject test of the ITBS (either reading, language arts, mathematics, science, or social studies) and the entire MMAT were administered to each group. Students in the tenth grade sample were randomly assigned to one of four groups. One subject test of the TAP (either reading, mathematics, science, or social studies) and the entire MMAT were administered to each group. The specific subtests making up each ITBS/TAP subject test are listed in Appendix B.

Counterbalanced administrations were not systematically incorporated into the procedure because of logistical constraints. It is possible that some degree of counterbalancing resulted even without such a provision.

Students who took only one of the two corresponding MMAT and ITBS/TAP subject tests were eliminated from the equating data base. The number of students taking each pair of corresponding subject tests at each grade is presented in Table 1. These groups exceed the minimum size of 400 examinees per test recommended by Brennan and Kolen (1987).

[Insert Table 1 about here]

Equipercentile Equating of Raw Scores

Raw scores on the corresponding MMAT and ITBS/TAP subject tests were equated using the equipercentile method (Angoff, 1984). An MMAT raw score was considered equivalent to the ITBS/TAP raw score that had the closest cumulative frequency in the sample. In the interest of brevity, tables of equivalent scores are not presented in this paper but are available from the authors upon request.

Equated ITBS/TAP raw scores were converted to national percentile ranks using conversion tables (Hieronymus and Hoover, 1986b; Scannell, 1986b). These data were then used to estimate percentile ranks for all other students who took the MMAT in the spring of 1987. As a result, each examinee's raw score on an MMAT subject test was used to estimate his/her national percentile rank in that subject.

Characteristics of MMAT and ITBS/TAP Subject Tests

Tables 2 through 6 present test length (number of items), mean, standard deviation, index of item difficulty (mean "p" value), and estimate of reliability (Kuder-Richardson 20 or 21) for the raw score distributions of corresponding MMAT and ITBS/TAP subject tests.

[Insert Tables 2 through 6 about here]

As would be expected, the raw score distributions of corresponding tests do not exhibit the identical properties called for by Angoff (1984). For example, differences in mean "p" values range from .01 at grades 6 and 10 mathematics to .22 at grades 3 and 8 social studies.

A number of pairs demonstrate strikingly similar characteristics, especially mathematics at grades 6, 8, and 10 and science at grade 10. Note in particular the similarities in item difficulty and reliability in corresponding subject tests of different lengths, such as mathematics and science at grade 10

Figures 1 through 19 graphically show the equating study raw score distributions of corresponding MMAT and ITBS/TAP subject tests. The shapes of the MMAT distributions vary. A few approximate symmetry, such as language arts at grade 8 and mathematics at grade 10, while skewness is apparent in reading at all four grades. Most of the ITBS/TAP distributions are, as would be expected, symmetrical. Several MMAT and ITBS/TAP and corresponding subject tests have remarkably similar distributions, such as mathematics and language arts at grade 6, language arts at grade 8, and mathematics at grade 10.

[Insert Figures 1 through 19 about here]

Scatterplots depicting the relationship of corresponding MMAT and ITBS/TAP subject test raw scores are shown in Figures 20 through 38. Some relationships appear to be linear, such as mathematics at grade 6 and language arts at grade 8. Most, however, are curvilinear. The scatterplots, as well as the graphs, indicate ceiling effects on several MMAT subject tests (e.g., reading at grades 3 and 6 and language arts at grade 3). Floor effects are not apparent from the data.

[Insert Figures 20 through 38 about here]

Pearson product moment coefficients relating corresponding MMAT and ITBS/TAP subject tests are presented in Table 7. These correlations range from .713 to .870; all exceed the Title I minimum for ORT-NRT equating of .60. In general, the correlations are relatively stable across subjects and across grades. The correlations for Chapter I program subjects--reading, language arts, and mathematics--are all quite similar. The correlations for mathematics at grades 6 and 8 show a slightly higher relationship than those for other subjects. The lowest correlation is between the science tests at grade 3.

[Insert Table 7 about here]

It is important to keep in mind that a test cannot correlate more highly with any other score than it correlates with its own true score (Allen and Yen, 1979), so the reliabilities of two corresponding subject tests set the upper limit of their correlation coefficient.

Estimated Comparable Percentiles, Not Equated Percentiles

Equating is a term reserved for linking scores on two tests that measure the same psychological function (Angoff, 1984). As noted, corresponding ITBS/TAP and MMAT subject tests measure similar but not identical content and have similar but not identical psychometric properties. According to the *Standards for Educational and Psychological Testing* (APA, 1985) this type of conversion should not be regarded as yielding equated or interchangeable scores but rather as having been done to achieve comparability.

> The procedures used to achieve comparability may be the same as those used in test equating, but the strict requirements of test equating will not be satisfied and, therefore, the resulting scores should be called scaled or comparable rather than equated. (p.32)

Thus, norm-referenced scores derived from the MMAT are presented as estimated comparable national percentile ranks rather than as equated percentile ranks.

Practical Utility of Estimated Comparable Scores

Accuracy of Estimated Comparable National Percentile Ranks

It was not possible to cross validate the results of the 1987 equating study due to practical constraints, so the accuracy of the MMAT estimated comparable

national percentile ranks at either the individual student level or the aggregate level has not yet been investigated. This is a techn l issue of the Missouri application of the ORT ONLY MODEL which will dressed in future equatings. Nevertheless, several methodological facto undoubtedly contributed to the accuracy of the estimated comparable national percentile ranks:

the NRT chosen for equating is similar in content to the ORT,

corresponding NRT and ORT subject tests were equated using an adequate number of examinees,

corresponding NRT and ORT subject tests were equated using a single-group design,

corresponding NRT and ORT subject tests were equated using the most appropriate method for such data--equipercentile equating, and

the equating sample for each pair of corresponding NRT and ORT subject tests was representative of the population to whom the results were applied.

Moreover, preliminary analyses of the data show that:

corresponding NRT and ORT subject tests share similar psychometric properties, and

corresponding NRT and ORT subject tests have correlations that exceed Title I/Chapter I guidelines for ORT-NRT equating.

These factors represent strengths of this application and suggest that the estimated comparable national percentile ranks should be considered as having practical utility for specific purposes. A discussion of how they were reported and how they can be used, at the individual student and the aggregate level, follows.

Individual Student Scores

Reporting Estimated Comparable National Percentile Ranks

Estimated comparable percentile ranks were reported for students in grades 3, 6, and 8 in reading, language arts, mathematics, science, and social studies and for students in grade 10 in all subjects except language arts. A 1987 MMAT Individual Student Report in reading/language arts is shown in Appendix C. It lists the student's estimated comparable national percentile rank, but

emphasizes objective-referenced information--subject and cluster (a group of related Key Skills) scores and Key Skill mastery data. A student's comparable scores were also listed on his/her 1987 MMAT Student Score Report Label (an adhesive-backed label for a permanent record).

The estimated comparable percentile ranks were used to prepare a special report called the MMAT Chapter I Eligibility List. This report lists the students in a designated grade that are eligible for Chapter I services in one, two, or all three subjects: reading, mathematics, and language arts. Each student's estimated comparable national percentile rank and its corresponding normal curve equivalent is listed. Appendix D presents the Chapter I eligibility standards for each grade, and Appendix E is a 1987 MMAT Chapter I Eligibility List.

Using Comparable Scores for Chapter I Purposes

As stated previously, the ORT ONLY MODEL was implemented in Missouri primarily to enable districts to utilize MMAT results for Chapter I purposes. Thus, estimated comparable national percentile ranks in reading, language arts, and mathematics are used for determining eligibility for placement in respective Chapter I programs as well as for program evaluation.

The use of MMAT estimated comparable national percentile ranks for Chapter I purposes does not conflict with Angoff's guidelines regarding the use of comparable scores. Because these scores are obtained from equating two nonparallel tests, they imply a level of achievement that is relative (or comparable) from one test to another. For example, the performance of a student scoring at the 40th percentile on the MMAT is relative to that same level of achievement on the ITBS/TAP.

As previously mentioned, floor and ceiling effects can induce error into the equating process (Fishbein, 1978). Floor effects are not apparent in this ORT-NRT equating, although most students in Chapter I programs would not score so low that floor effects would be a problem. In fact, most eligibility cutoff scores (see Appendix D) are within the range of the score distribution where equating error is likely to be minimal. Ceiling effects are less likely to be a problem with respect to Chapter I applications than they are for other purposes, such as identifying academically talented and gifted students (which is discussed in the following section).

It is important to keep in mind that the decision to report individual student estimated comparable national percentile ranks was based on the need to provide a mechanism for minimizing testing time and cost at the local district level. The estimated comparable scores are, therefore, reported for the convenience of MMAT users in need of data for Chapter I applications. However, their validity for such purposes has not yet been empirically

determined; this will be the focus of further investigations of the Missouri procedure.

Using Comparable Scores for Gifted Education Program Purposes

There is likely to be more error in estimating comparable scores in the tails of the score distribution than in the middle range (Roudabush, 1975). Ceiling effects, one source of error in the upper tail, were apparent in the MMAT distributions. These effects suggested that the comparable scores should not be used to identify students for placement in gifted education programs. Subject scaled scores, derived using item response theory (see Appendix A), seemed to be much more appropriate for such a purpose. Consequently, MMAT users were provided with the state percentile ranks of subject scaled scores for the purpose of identifying academically talented and gifted students.

Using Comparable Scores as Substitutes for NRT Scores

Until their accuracy can be empirically determined, estimated comparable national percentile ranks should probably not be routinely used as substitutes for actual NRT scores. Teachers, counselors, and administrators were, therefore, discouraged from treating these scores as if they were equivalent to those resulting from administration of an NRT. This is in keeping with the distinction between equated and comparable scores given in the *Standards for Educational and Psychological Testing* (APA, 1985):

> To say that scores have been made comparable is a weaker claim than to say that they have been equated. Equated scores are meant to be interchangeable, whereas comparable scores are meant to be similar in a particular sense. (p. 32)

Because estimated comparable national percentile ranks were reported on 1987 Individual Student Reports, users tended to put more stock in them than was appropriate. The norm-referenced information will not be presented on the 1988 Individual Student Reports in an attempt to minimize misinterpretation. In 1988, estimated comparable national percentile ranks will only be reported on the Student Score Report Label and the Chapter I Eligibility List.

The decision to exclude NRT scores from Individual Student Reports will hopefully eliminate problems resulting from the different methods used to scale MMAT results. Estimated comparable national percentile ranks, as well as Key Skill mastery results, were obtained using raw scores, while the subject scaled scores were obtained using a two-parameter item response theory model (see Appendix A). The different scales caused some understandable confusion on the part of MMAT users, because it was possible for two students to achieve the same estimated comparable national percentile rank

but different scaled scores and different Key Skill mastery results in a particular subject.

Aggregate Scores

MMAT estimated comparable national percentile ranks were not reported at the aggregate level for the state sample or the population in 1987. Several districts computed median estimated comparable national percentile ranks in order to report aggregate level data to patrons. Aggregate level comparable scores could probably be used with confidence to assess the standing of groups of students relative to their national peers.

It will be another year before Chapter I program evaluation data can be analyzed, because districts are not required to collect it this academic year while making the transition from an NRT to the MMAT.

Summary and Conclusions

The ORT ONLY MODEL provides a mechanism for reporting norm-referenced information from an assessment instrument that emphasizes objective- or criterion-referenced test information. It eliminates redundant testing, thereby saving time and money. To truly maximize the efficiency of the ORT ONLY MODEL, norm-referenced information is needed at the individual student level. However, when NRT data are reported for individual students, several issues need to be considered: (a) the accuracy of the individual student scores, (b) the use of appropriate procedures in order to maximize score accuracy, and (c) the practical utility of individual student and aggregate scores.

The ORT ONLY MODEL is being implemented in Missouri to obtain norm- and objective-referenced information from the newly-developed statewide assessment. Equipercentile equating, using a single-group design, was used to obtain norm-referenced scores from the MMAT for the first time in 1987. Norm-referenced data, referred to as estimated comparable national percentile ranks, were primarily reported at the individual student level and for Chapter I purposes. Estimated comparable national percentile ranks will be obtained annually, using the same method and design, to improve the accuracy of the NRT data.

While it was not possible to cross validate the 1987 MMAT estimated comparable national percentile ranks, preliminary analyses showed that the corresponding MMAT and ITBS/TAP subject tests from which they were obtained are similar in terms of content and statistical properties. Moreover, correlation coefficients of corresponding subject tests were at acceptable levels

and the equating sample was representative of the population to whom the results were applied.

Thus preliminary data analyses, as well as the use of appropriate equating procedures, provide support for using the individual student estimated comparable scores for Chapter I purposes. Because of ceiling effects, individual student estimated comparable scores should not be used for identifying academically talented students. They should also not be viewed as equivalent to scores obtained from an NRT.

Applied measurement research is frequently conducted in less than ideal circumstances. The initial stage of this ongoing investigation of the ORT ONLY MODEL was conducted in the context of a newly implemented statewide testing program. Its shortcomings and merits will hopefully be judged accordingly.

Recommendations for Further Research

There is much to be learned about using the ORT ONLY MODEL, both in terms of whether it is a viable model and in terms of how to make certain that, if it is used, it yields valid results. Future investigations should focus on the following:

> the worth of the ORT ONLY MODEL relative to the other three models,
>
> the appropriateness of equipercentile equating for obtaining comparable scores,
>
> the effects of content and test level on the equating results,
>
> the accuracy of comparable scores at the individual student level,
>
> the accuracy of student level comparable scores in the low, middle, and high ranges of the distribution,
>
> the validity of specific uses of comparable scores,
>
> the effects of annual recalibration on the accuracy of comparable scores, and
>
> the effects of instruction and, as a result, increasingly skewed ORT data, on the accuracy of comparable scores.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth, Inc.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Bauer, E. A. (1979, April). How minimal is minimal? Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED 177 228)

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. Applied Psychological Measurement, 11, 279-290.

Bunch, M. B. (1982, March). Using non-normed tests in Title I evaluation. Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC Document Reproduction Service No. ED 220 492)

Crane, L. R., Prapuclenis, P. G., Rice, W. K., & Perlman, C. (1981). The effect of different equating methods on Title I evaluation model A2 NCE gain estimates (Contract No. 300-79-0485). Evanston, IL: Educational Testing Service.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Echternacht, G. (Ed.). (1980). Measurement aspects of Title I evaluations. San Francisco: Jossey-Bass Inc.

Fishbein, R. L. (1978, March). The use of non-normed tests in the ESEA Title I evaluation and reporting system: Some technical and policy issues. Paper presented at the annual meeting of the American Educational Research Association, Toronto. (ERIC Document Reproduction Service No. ED 159 176)

Hieronymus, A. N., & Hoover, H. D. (1986a). Iowa tests of basic skills. Chicago: Riverside Publishing Co.

Hieronymus, A. N., & Hoover, H. D. (1986b). National norms for forms G & H levels 5-14. Chicago: Riverside Publishing Co.

Holmes, S. E. (1980, January). ESEA Title I evaluation and reporting refinement: The Title I linking project (Final Report). Oregon Department of Education.

Keene, J. M., & Holmes, S. E. (1987, April). Obtaining norm-referenced test information for local objective-referenced tests: Issues and challenges. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Linn, R. L. (1978, March). The validity of inferences based on the proposed Title I evaluation models. Paper presented at the annual meeting of the American Educational Research Association, Toronto. (ERIC Document Reproduction Service No. ED 156 696)

Missouri Department of Elementary and Secondary Education. (1986). Testing standards for Missouri public schools. Jefferson City, MO: Author.

Osterlind, S. J. (1987). Missouri mastery and achievement tests. Jefferson City, MO: Missouri Department of Elementary and Secondary Education.

Riverside Publishing Company. (1986a). Preliminary technical summary. Chicago: Author.

Riverside Publishing Co. (1986b). Iowa tests of basic skills/tests of achievement and proficiency correlated to key skills for Missouri schools. Chicago: Author.

Roudabush, G. E. (1975, April). Estimating normative scores from a criterion-referenced test. Paper presented at the annual meeting of the American Educational Research Association in Washington, DC. (ERIC Document Reproduction Service No. ED 106 352)

Scannell, D. P. (1986a). Tests of achievement and proficiency. Chicago: Riverside Publishing Co.

Scannell, D. P. (1986b). Tests of achievement and proficiency spring national percentile ranks for pupils. Chicago: Riverside Publishing Co.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. Applied Psychological Measurement, 10 (3), 303-317.

Storlie, T. R. (1979, April). An empirical comparison of Title I NCE gains estimated with model A1 and with model A2. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 200 646)

Texas Education Agency. (1986). Report on providing national comparative data on the TEAMS test. Austin: Author.

Table 1

Number of Examinees Taking Corresponding MMAT and ITBS/TAP Subject Tests

	Grade			
Subject tests	3	6	8	10
Reading	1511	1327	1256	1463
Language Arts	1439	1245	1202	--
Mathematics	958	1258	1406	1466
Science	1463	1082	1250	1269
Social Studies	1171	1230	1768	1523

Table 2

Properties of Corresponding Reading Tests

	Grade							
	3		6		8		10	
	ITBS	MMAT	ITBS	MMAT	ITBS	MMAT	TAP	MMAT
No. of Items	44	52	56	52	58	60	58	60
Mean	28.06	40.86	33.77	39.07	33.09	41.95	39.52	44.87
S.D.	9.18	9.79	10.43	8.92	10.97	10.22	10.94	9.56
Mean "P"	.64	.79	.61	.75	.57	.70	.69	.75
KR-20	.902	.946	.909	.928	.918	.942	.915	.940

Note: These data were computed on the raw score distributions of the equating sample.

Table 3

Properties of Corresponding Language Arts Tests

	Grade 3		Grade 6		Grade 8		Grade 10	
	ITBS	MMAT	ITBS	MMAT	ITBS	MMAT	TAP	MMAT
No. of Items	119	40	141	48	148	56	--	--
Mean	76.11	29.84	85.55	27.65	77.20	34.62	--	--
S.D.	20.29	7.36	23.09	8.02	22.03	10.30	--	--
Mean "P"	.64	.75	.61	.58	.52	.62	--	--
KR-21	.939	.877	.944	.834	.930	.887	--	--

Note: These data were computed on the raw score distributions of the equating sample.

Table 4

Properties of Corresponding Mathematics Tests

	Grade							
	3		6		8		10	
	ITBS	MMAT	ITBS	MMAT	ITBS	MMAT	TAP	MMAT
No. of Items	86	68	109	104	117	100	48	92
Mean	60.04	54.26	68.73	67.27	72.28	59.74	27.59	53.27
S.D.	14.03	11.04	18.70	17.29	21.57	18.01	8.80	16.47
Mean "P"	.69	.78	.64	.65	.62	.59	.60	.59
KR-20	.792	.922	.862	.946	.901	.939	.913	.933

Note: These data were computed on the raw score distributions of the equating sample.

Table 5

Properties of Corresponding Science Tests

	Grade							
	3		6		8		10	
	ITBS	MMAT	ITBS	MMAT	ITBS	MMAT	TAP	MMAT
No. of Items	38	64	43	92	45	72	54	80
Mean	20.37	44.69	22.41	51.08	22.41	39.67	27.73	38.07
S.D.	5.59	9.46	6.50	12.69	6.44	10.05	8.00	10.18
Mean "P"	.53	.70	.49	.56	.49	.56	.59	.53
KR-20	.764	.889	.820	.891	.814	.862	.854	.846

Note: These data were computed on the raw score distributions of the equating sample.

Table 6

Properties of Corresponding Social Studies Tests

	Grade							
	3		6		8		10	
	ITBS	MMAT	ITBS	MMAT	ITBS	MMAT	TAP	MMAT
No. of Items	38	56	43	84	45	72	62	100
Mean	18.63	37.0	21.96	58.361	22.13	49.15	43.22	67.76
S.D.	5.32	9.97	6.84	14.47	6.68	12.83	10.52	17.49
Mean "P"	.47	.69	.49	.70	.49	.71	.71	.65
KR-20	.756	.909	.821	.935	.856	.932	.914	.950

Note: These data were computed on the raw score distributions of the equating sample.

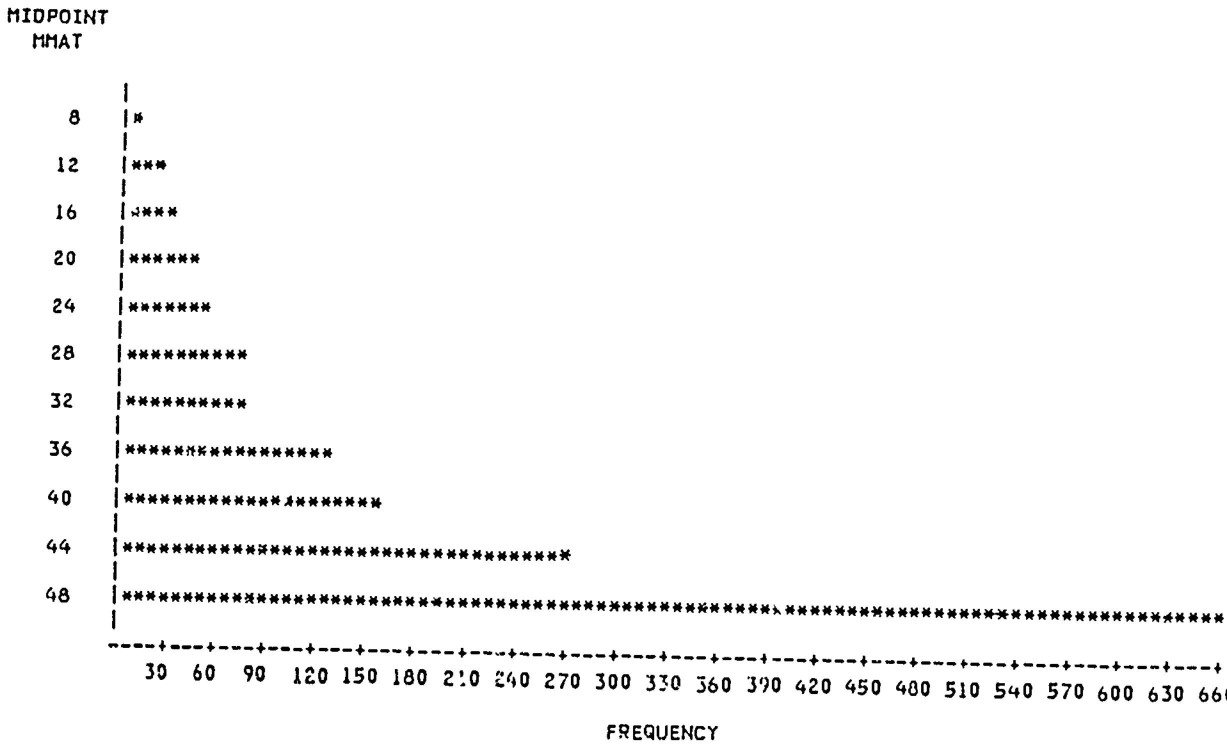
Table 7

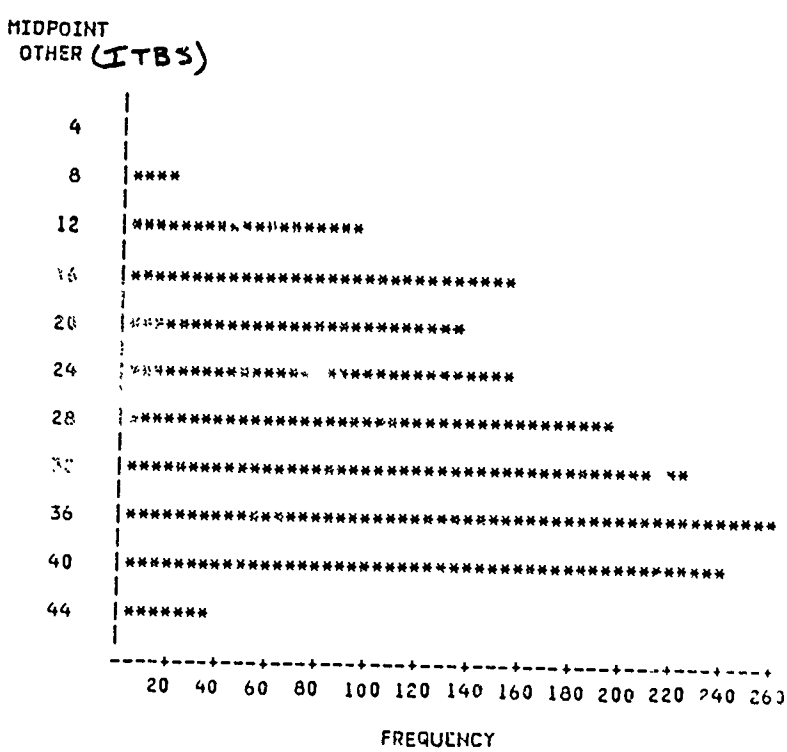
Pearson Product Moment Coefficients for Corresponding ITBS/TAP and MMAT Subject Tests

	Grade			
Subject tests	3	6	8	10
Reading	.799	.786	.808	.756
Language Arts	.799	.781	.792	--
Mathematics	.809	.860	.870	.856
Science	.713	.809	.743	.786
Social Studies	.759	.800	.789	.843

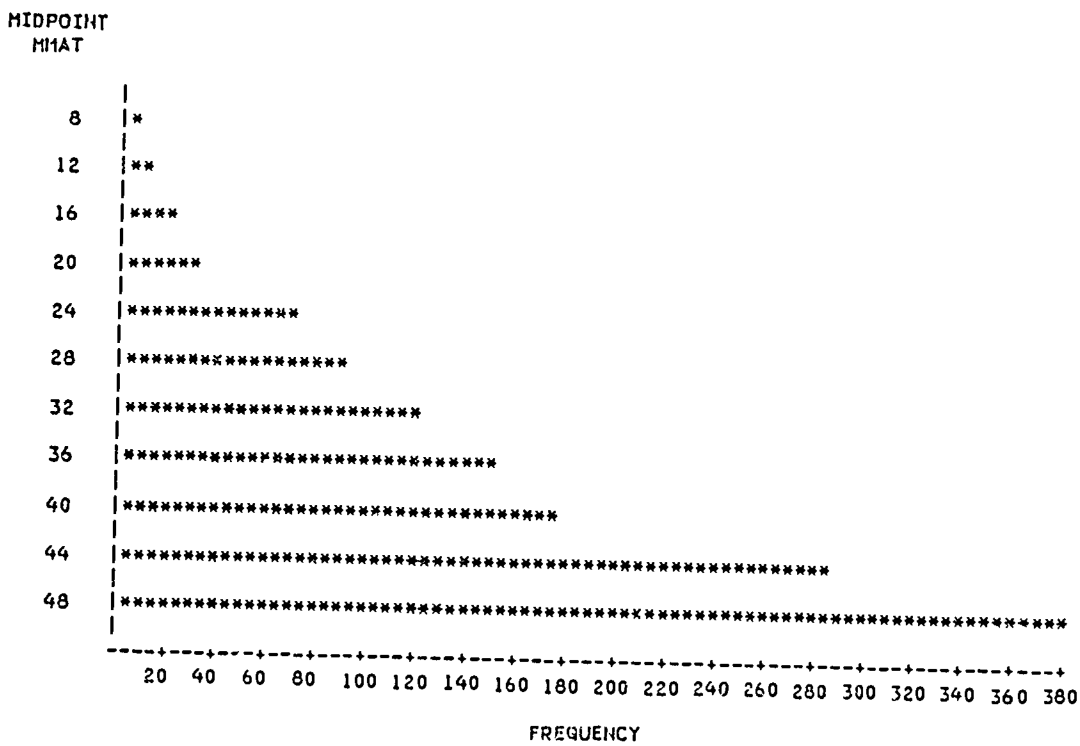
Figures 1 through 19

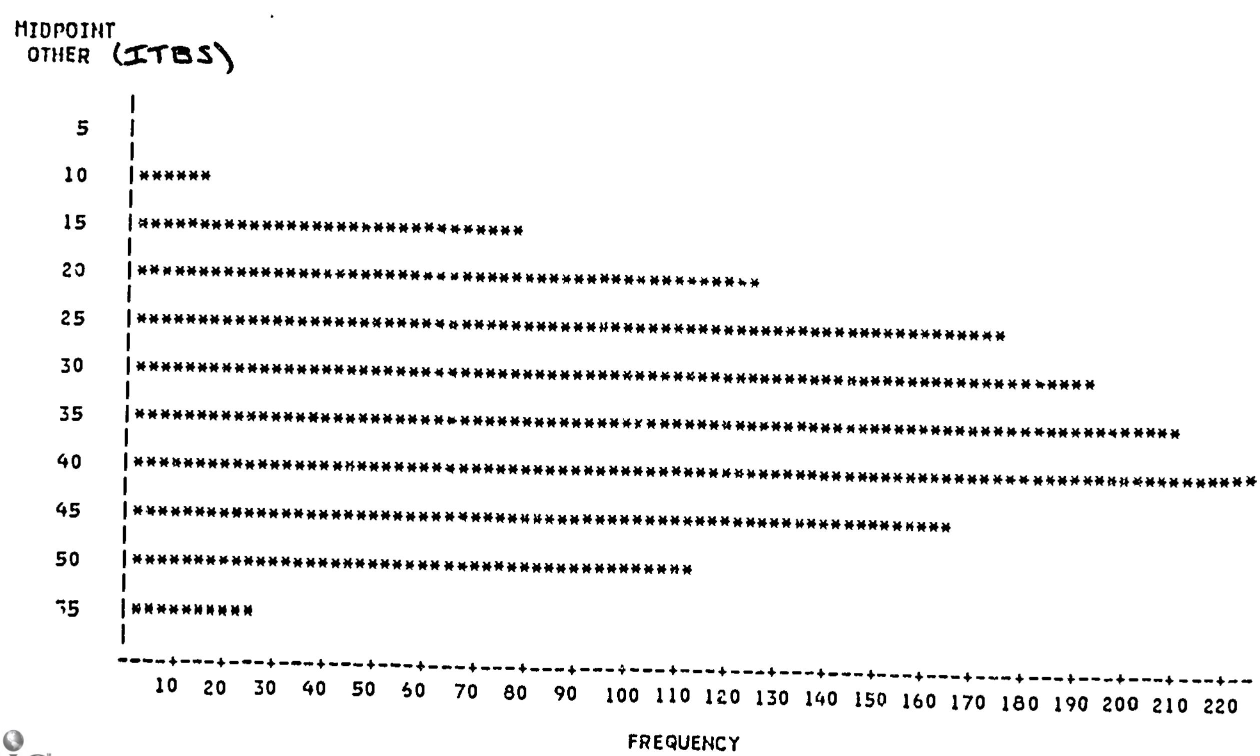
Raw Score Distributions of Corresponding MMAT and ITBS/TAP Subject Tests


MIDPOINT
MMAT
8
12
16
20
24
28
32
36
40
44
48
30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570 600 630 660
FREQUENCY

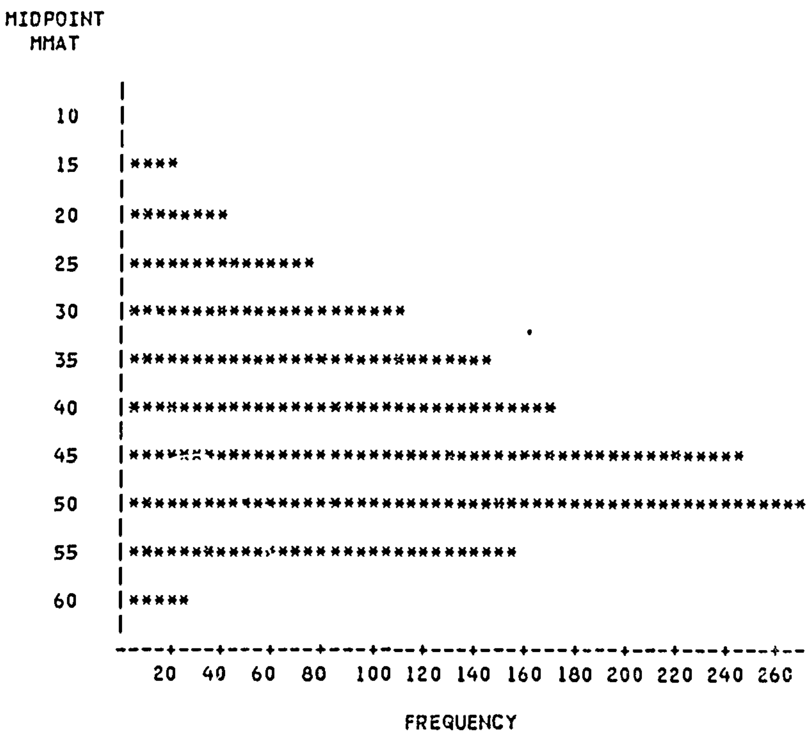

MIDPOINT
OTHER (ITBS)
4
8
12
16
20
24
28
32
36
40
44
20 40 60 80 100 120 140 160 180 200 220 240 260
FREQUENCY

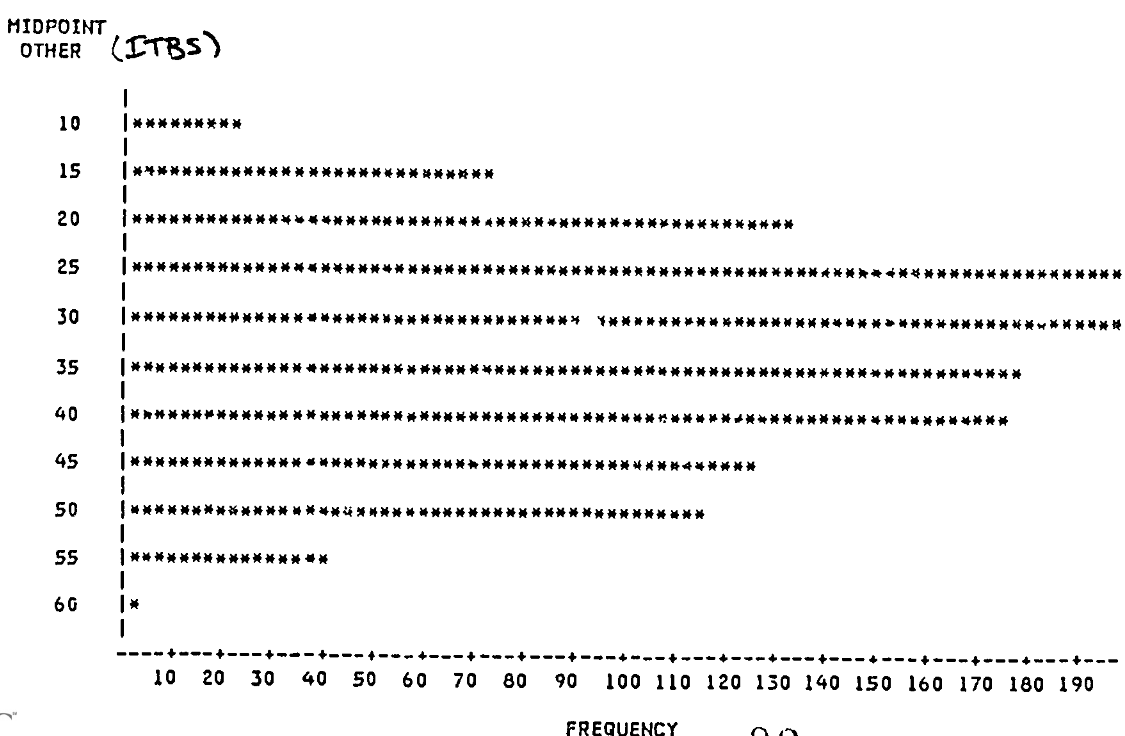
Grade 6 Reading

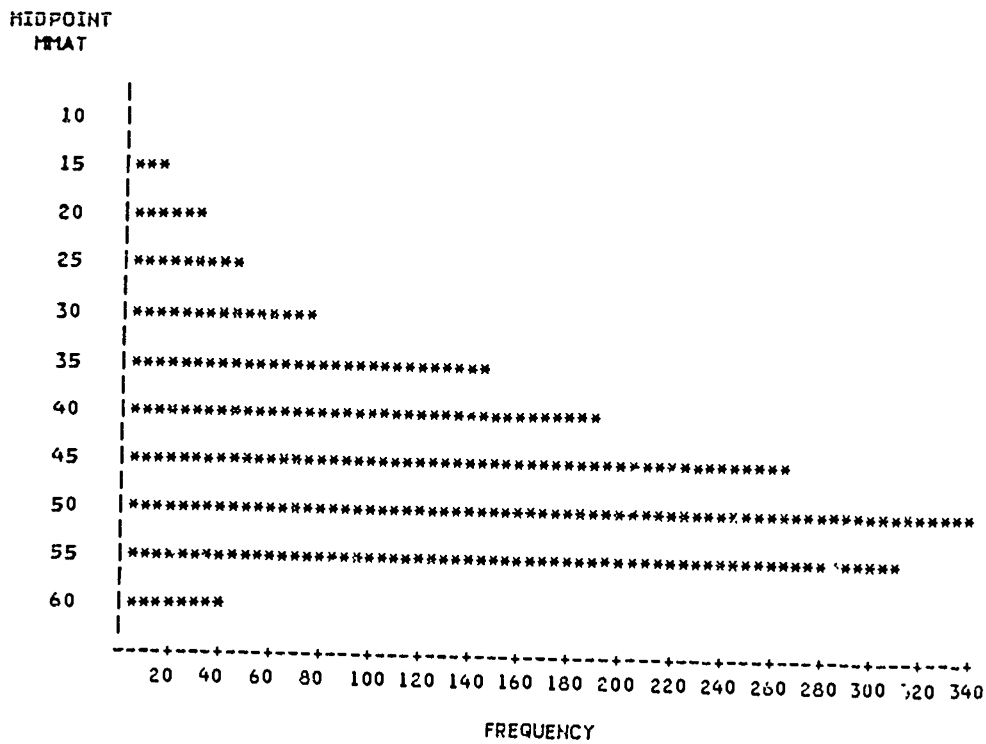


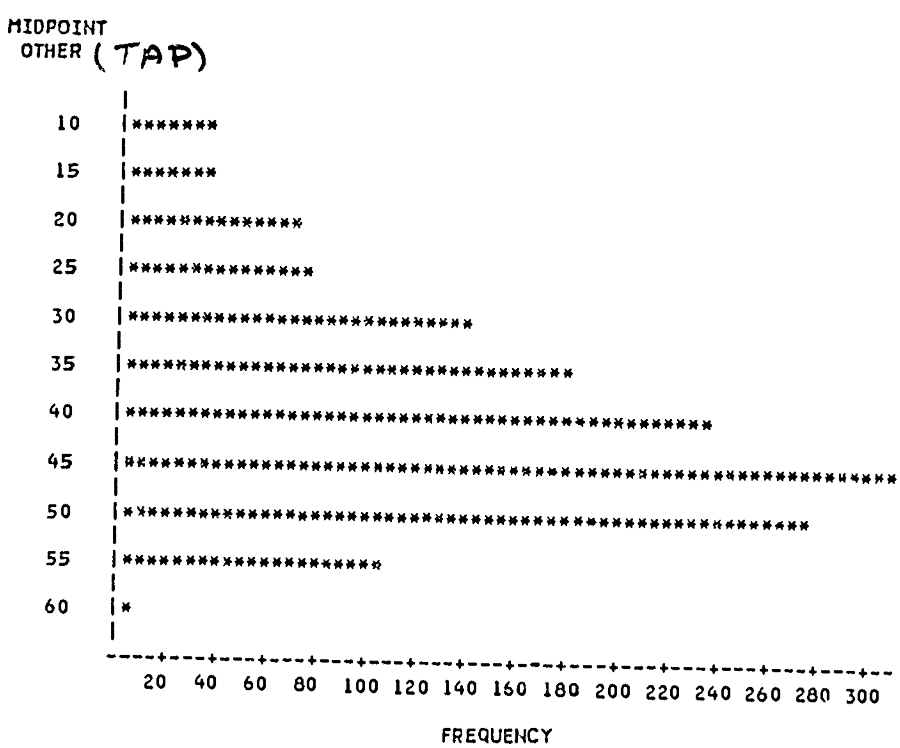


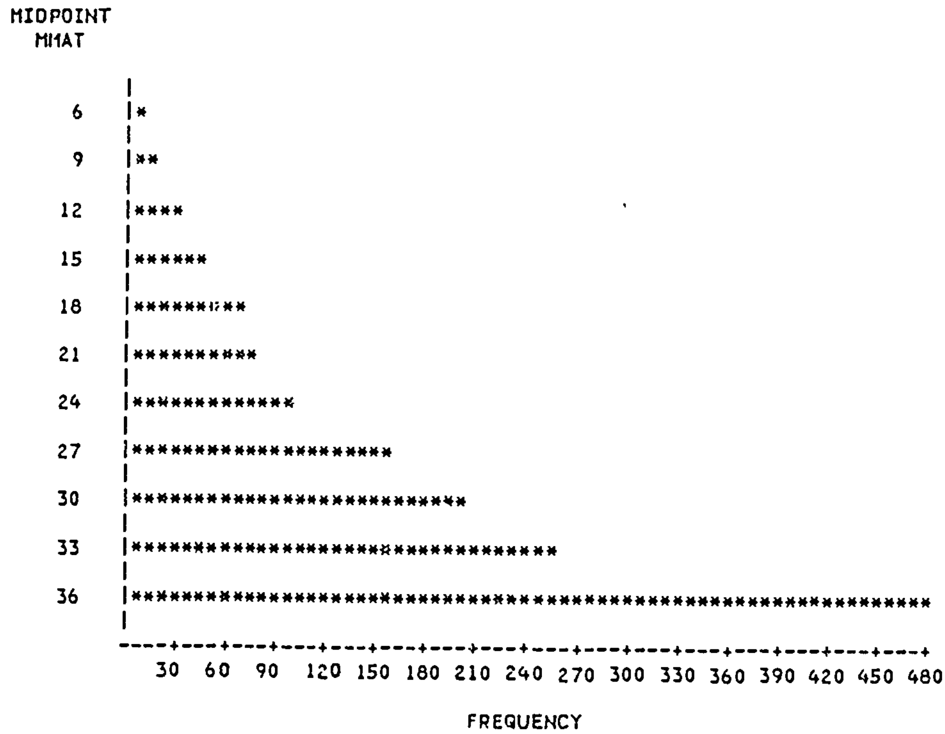
Grade 8 Reading

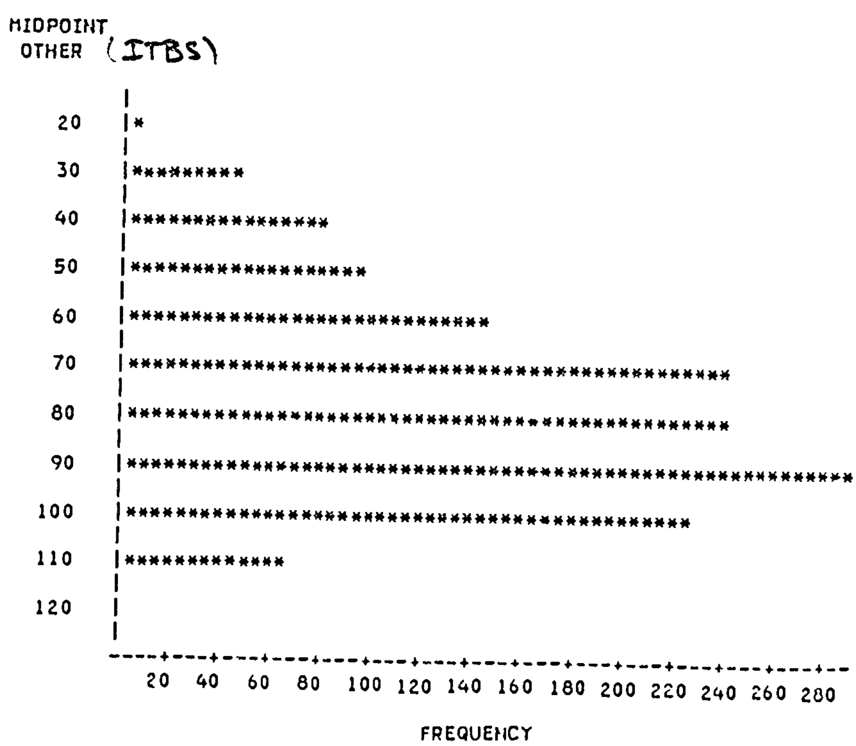




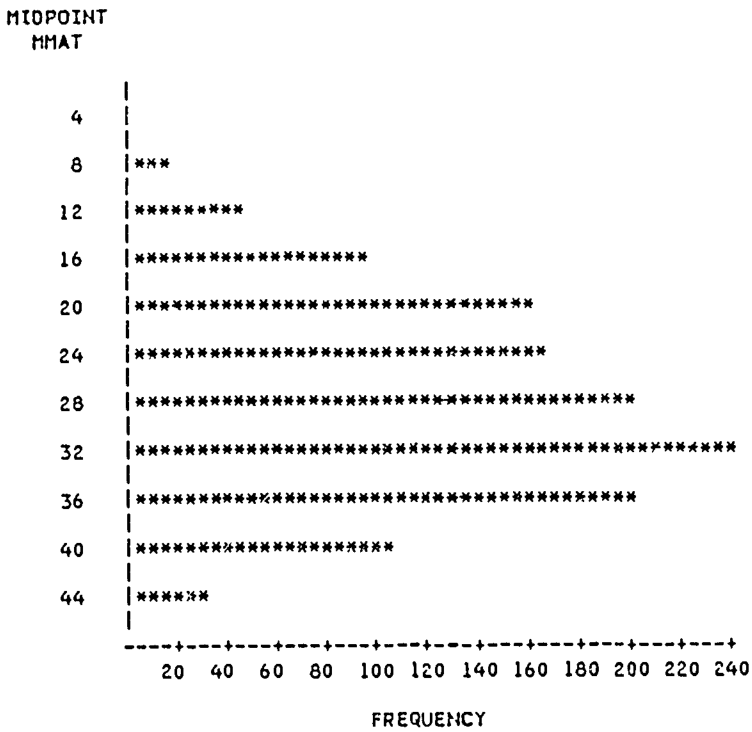

MIDPOINT
MMAT
10
15
20
25
30
35
40
45
50
55
60
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340
FREQUENCY

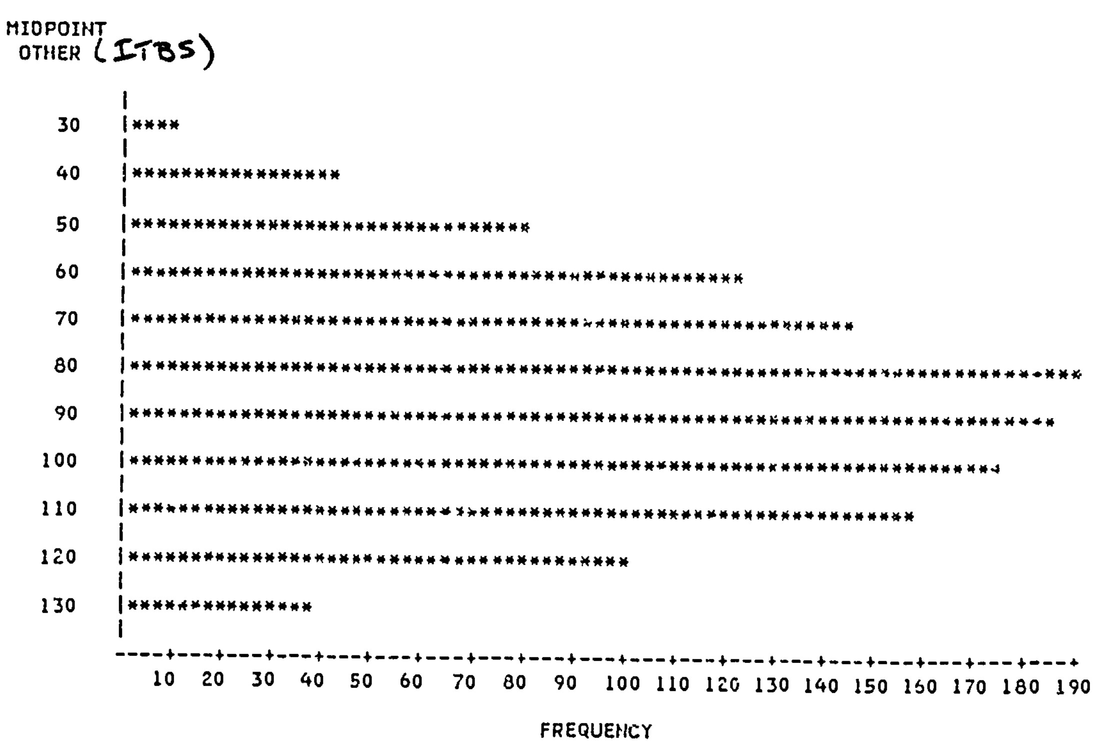

MIDPOINT
OTHER (TAP)
10
15
20
25
30
35
40
45
50
55
60
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
FREQUENCY

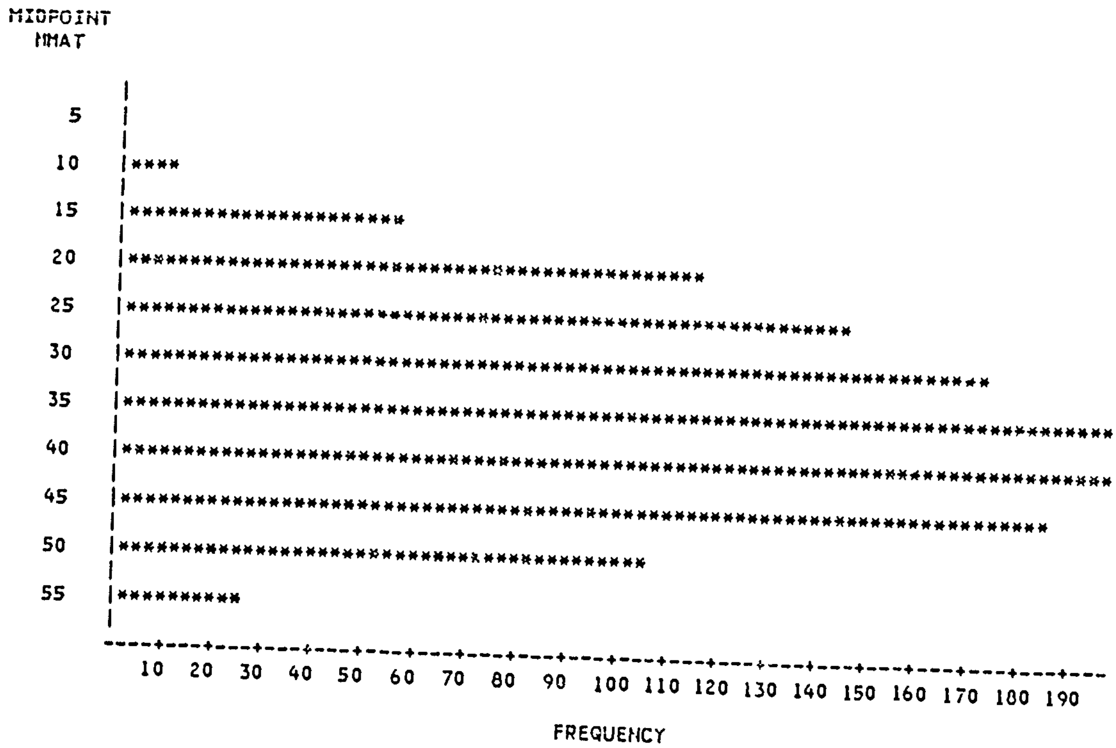

MIDPOINT
MMAT
6
9
12
15
18
21
24
27
30
33
36
30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480
FREQUENCY

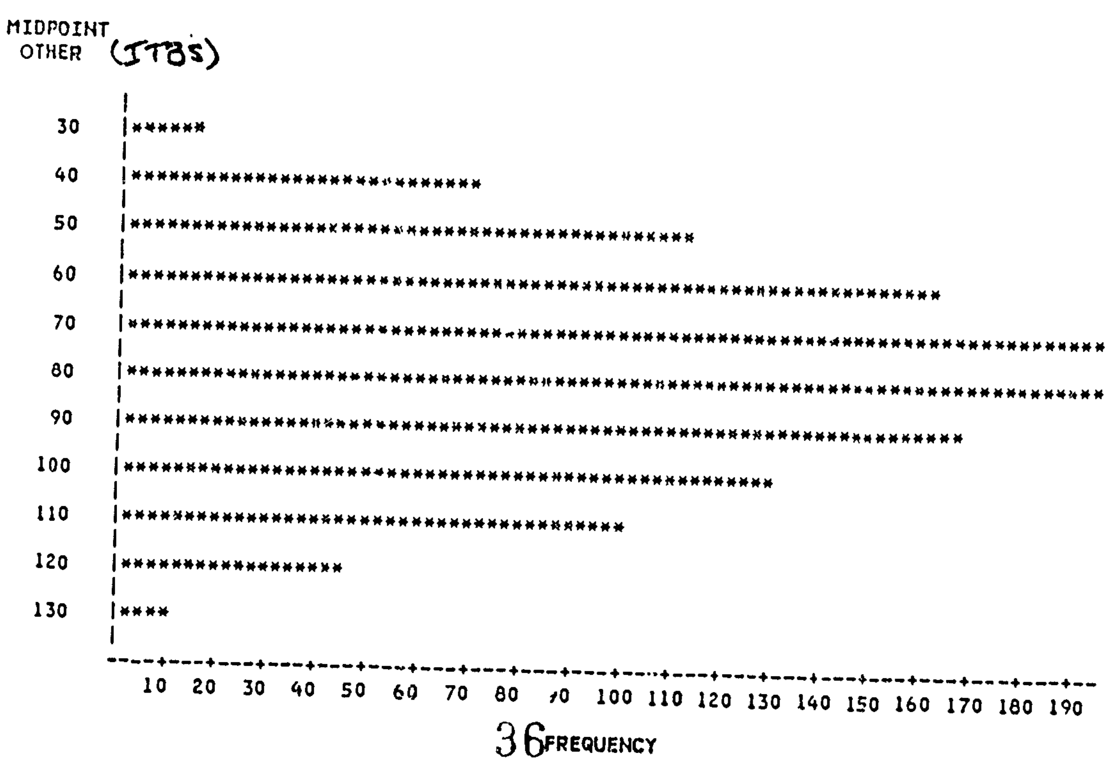

MIDPOINT
OTHER (ITBS)
20
30
40
50
60
70
80
90
100
110
120
20 40 60 80 100 120 140 160 180 200 220 240 260 280
FREQUENCY

Grade 6 Language Arts

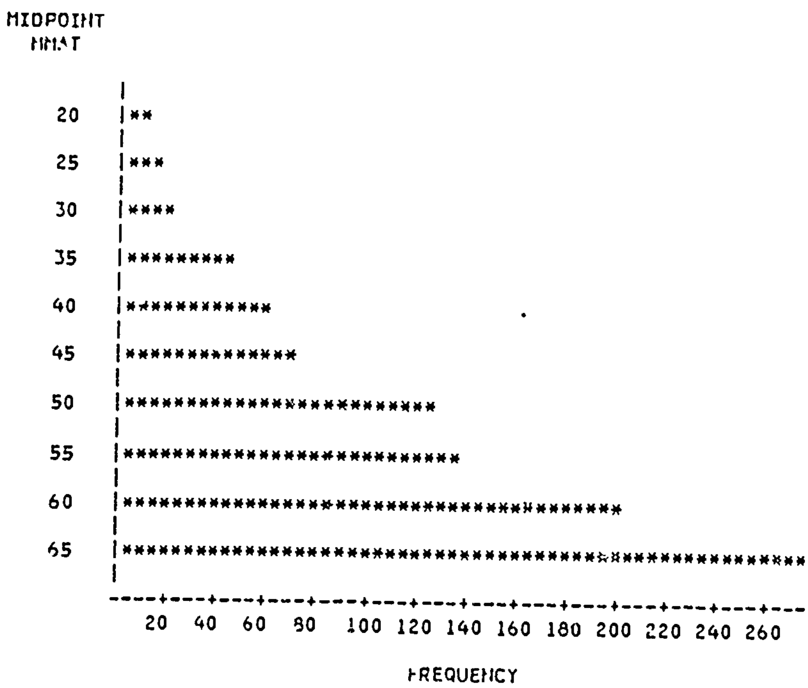


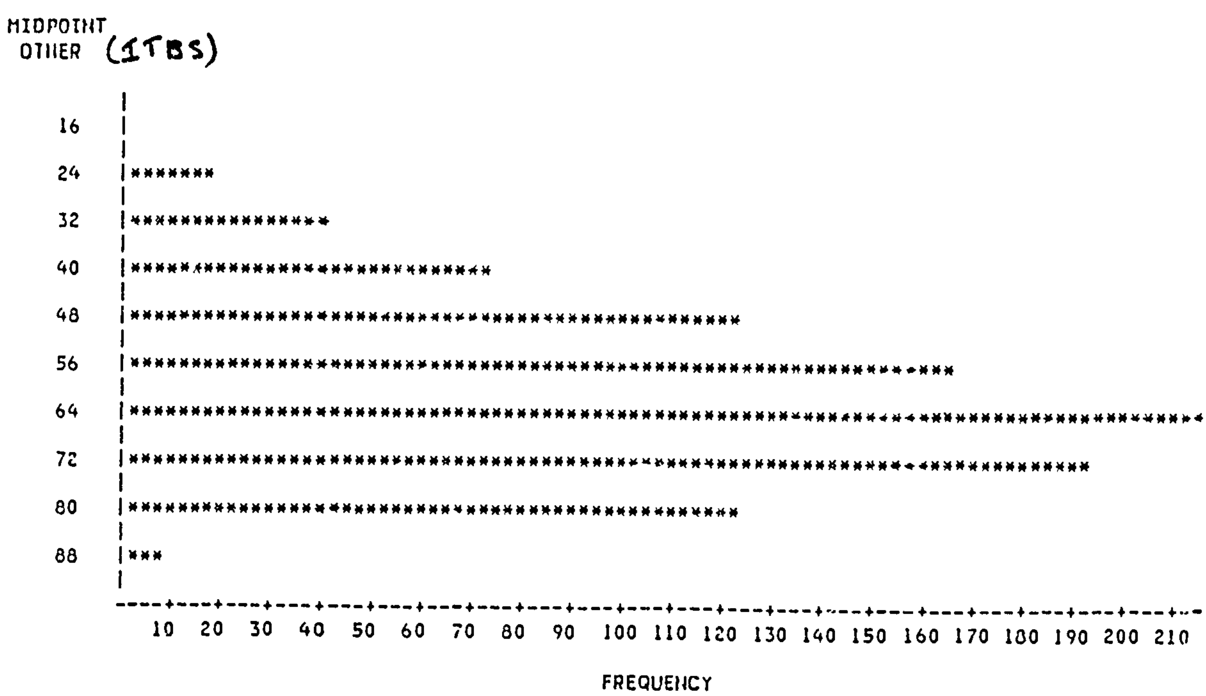



MIDPOINT
MMAT
5
10
15
20
25
30
35
40
45
50
55
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
FREQUENCY

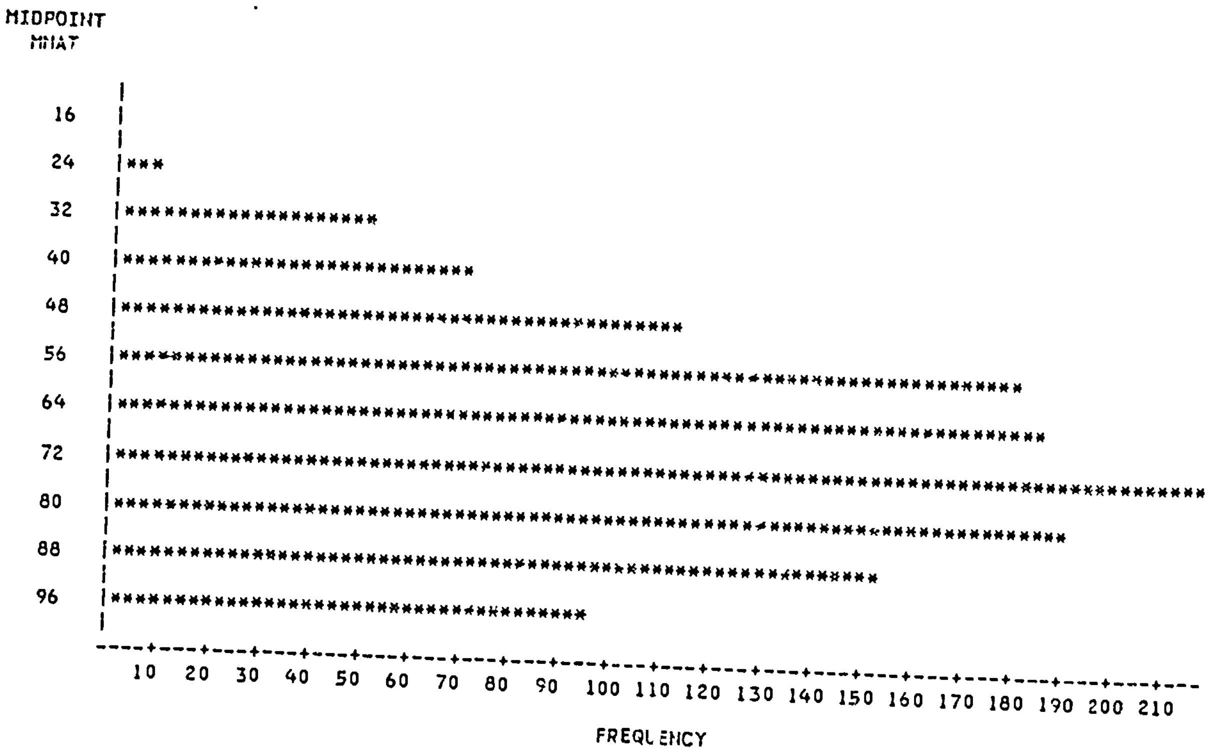

MIDPOINT
OTHER (ITBS)
30
40
50
60
70
80
90
100
110
120
130
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
FREQUENCY

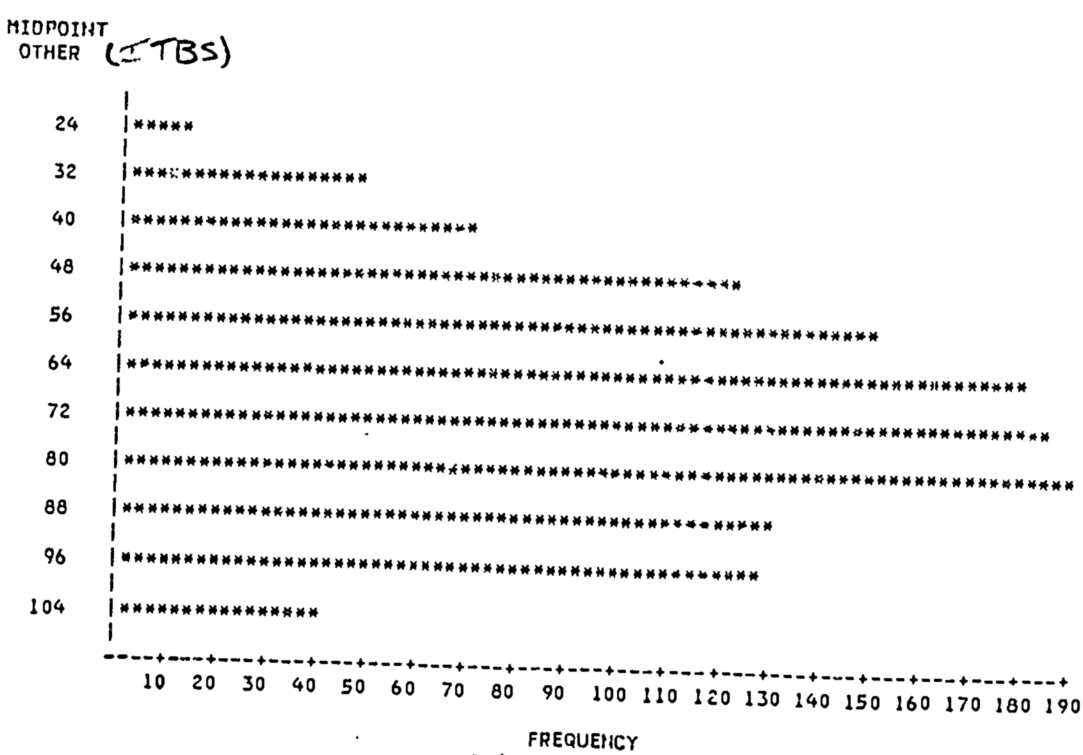
Grade 3 Mathematics

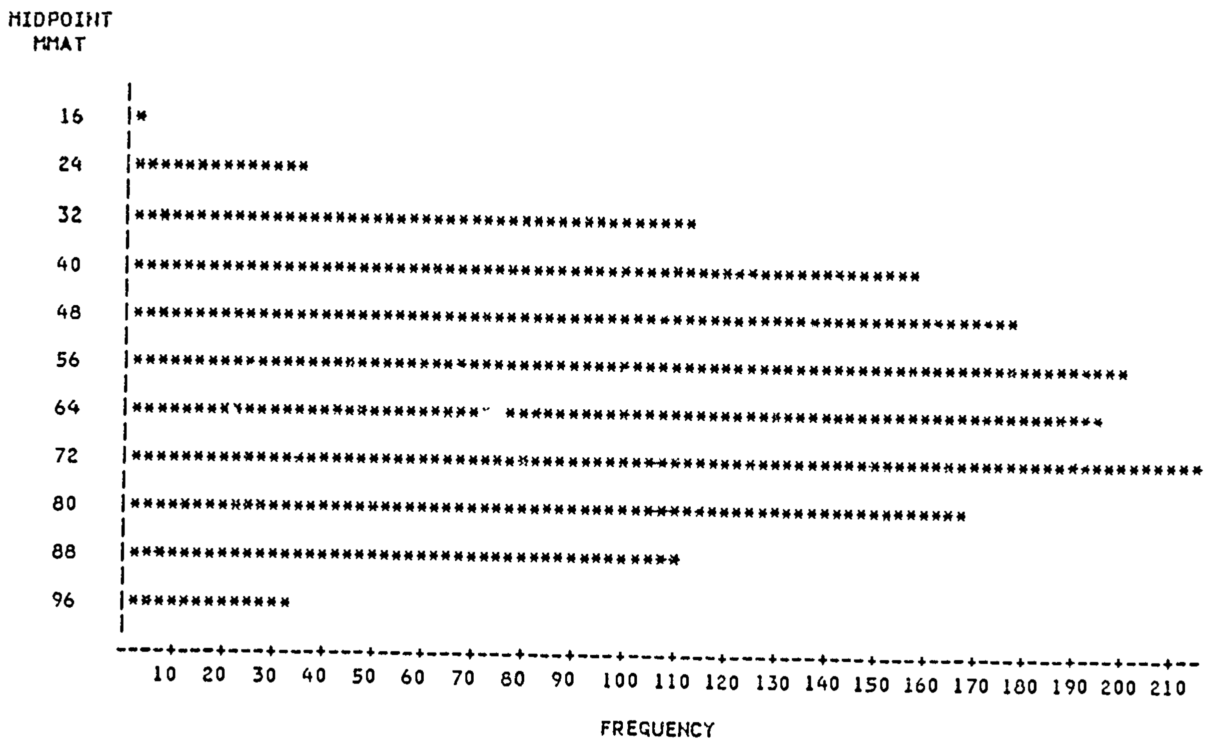


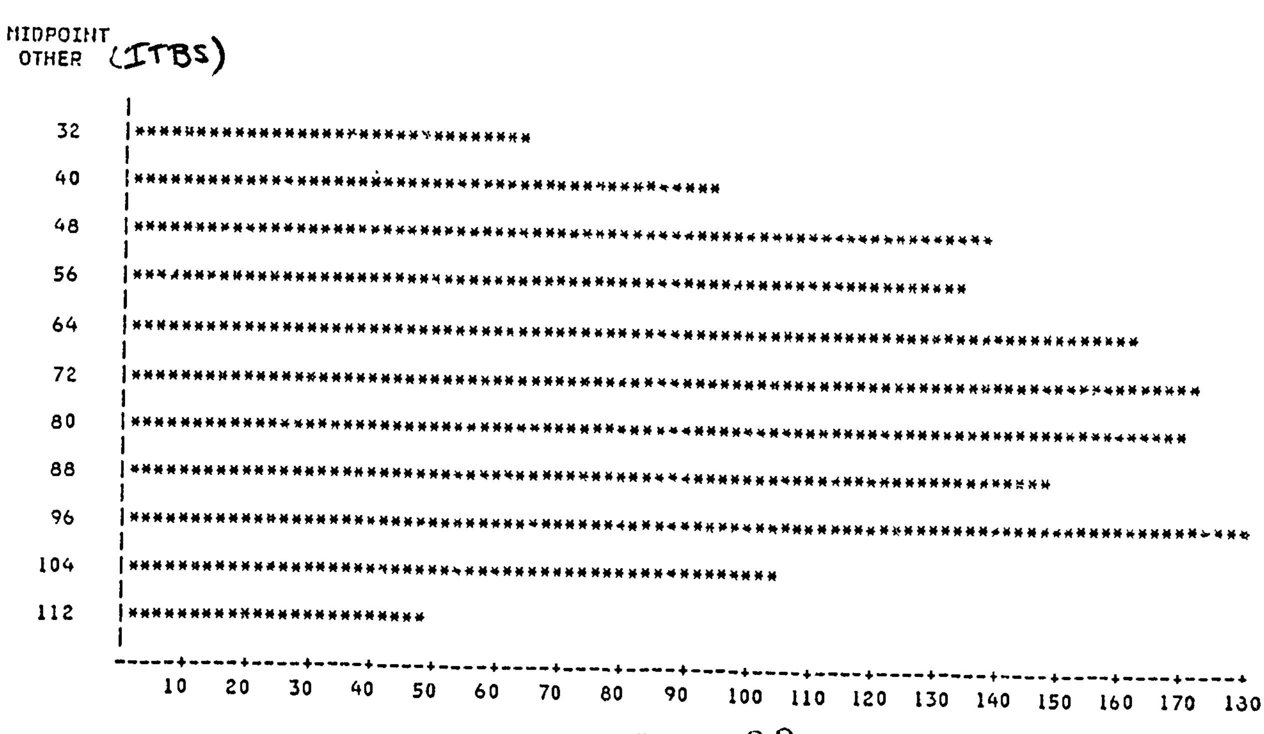


Grade 6 Mathematics

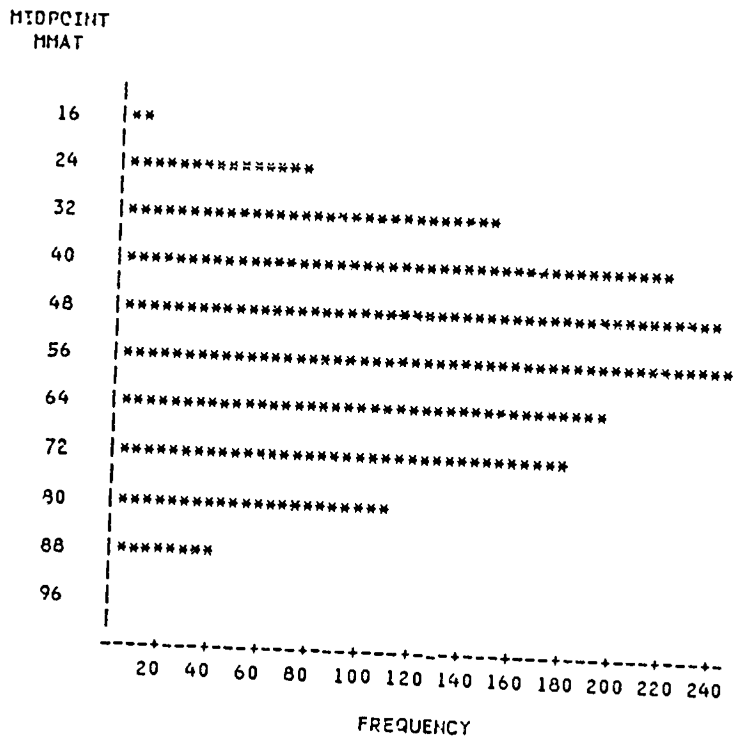


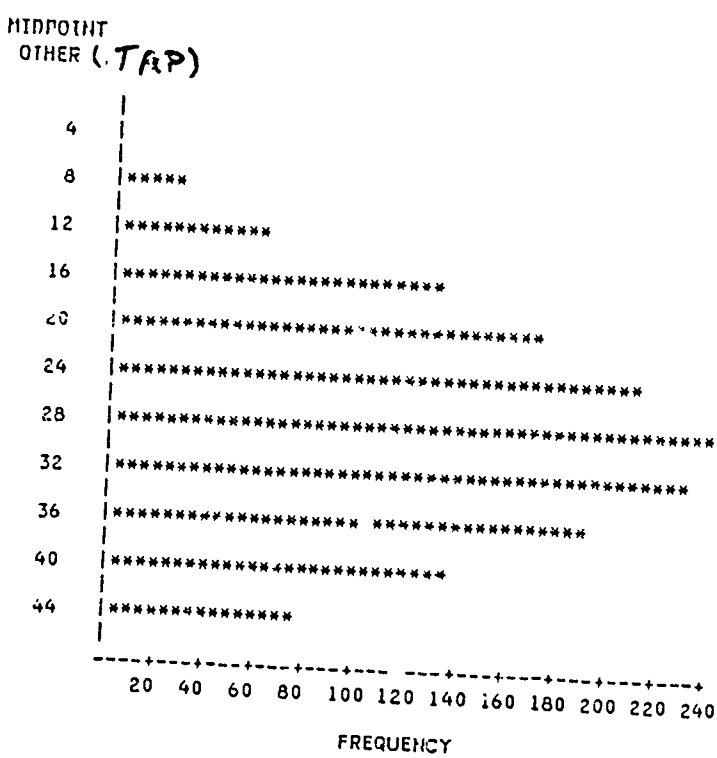



MIDPOINT
MMAT
16
24
32
40
48
56
64
72
80
88
96
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210
FREQUENCY

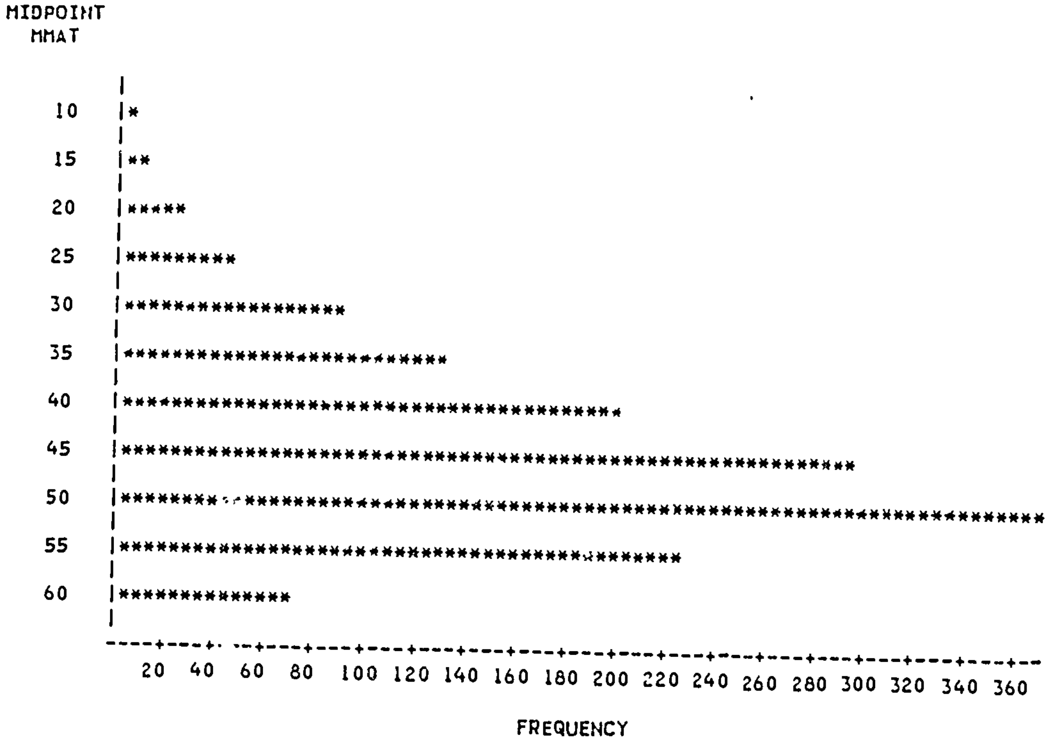

MIDPOINT
OTHER (ITBS)
32
40
48
56
64
72
80
88
96
104
112
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 130
FREQUENCY

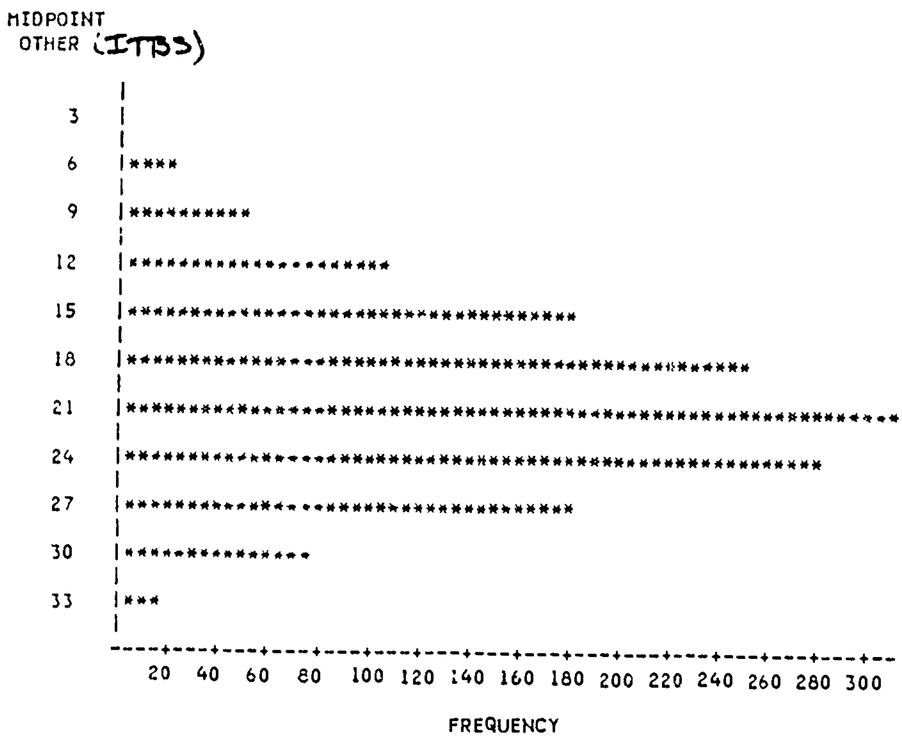
Grade 10 Mathematics



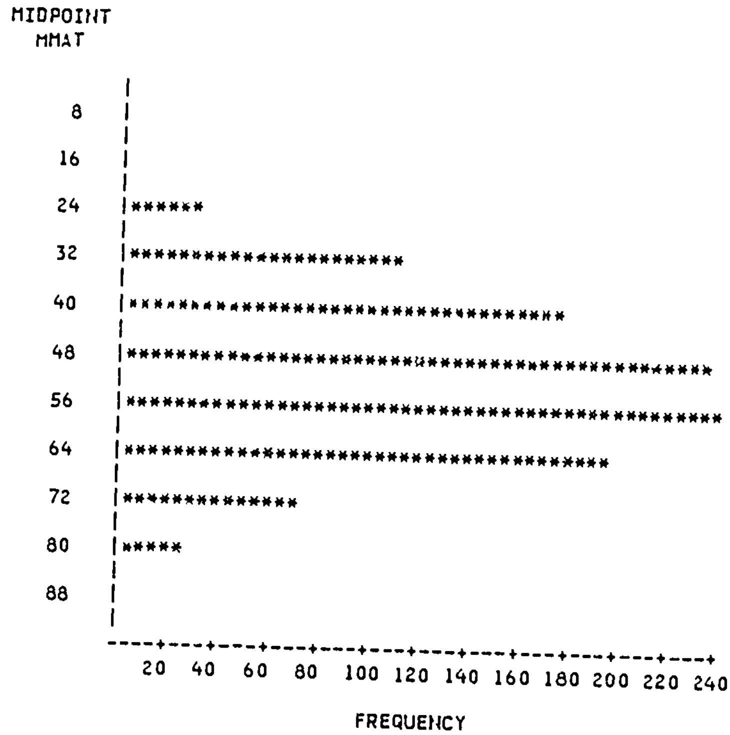


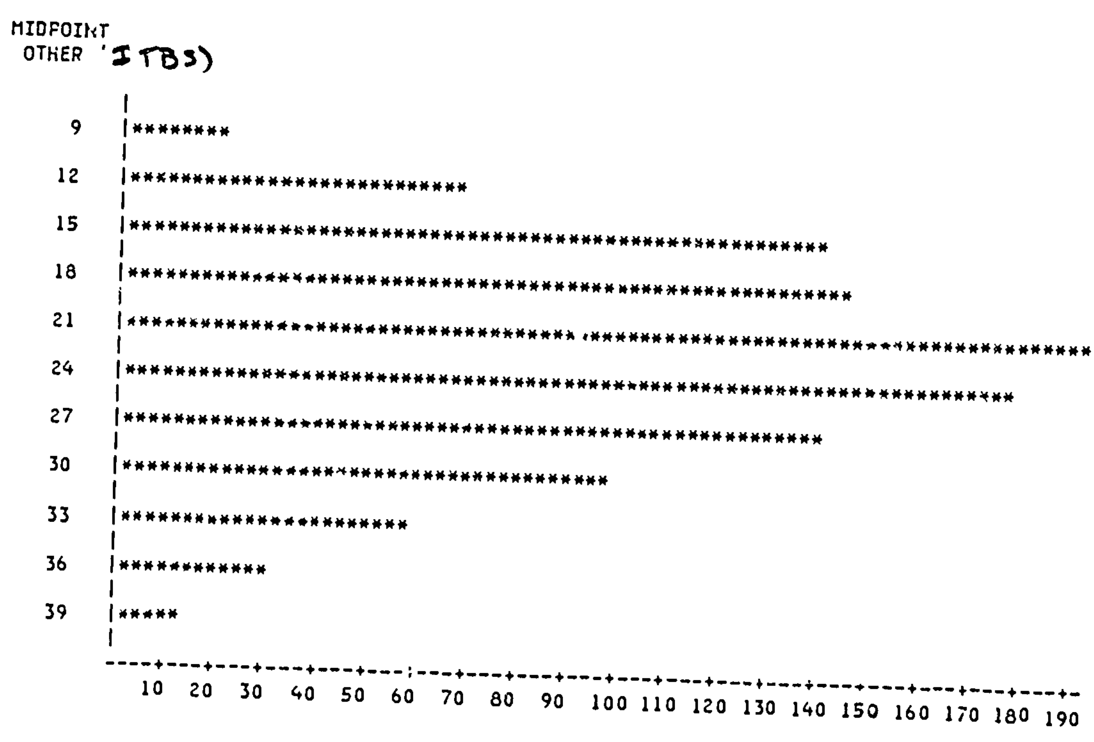
Grade 3 Science



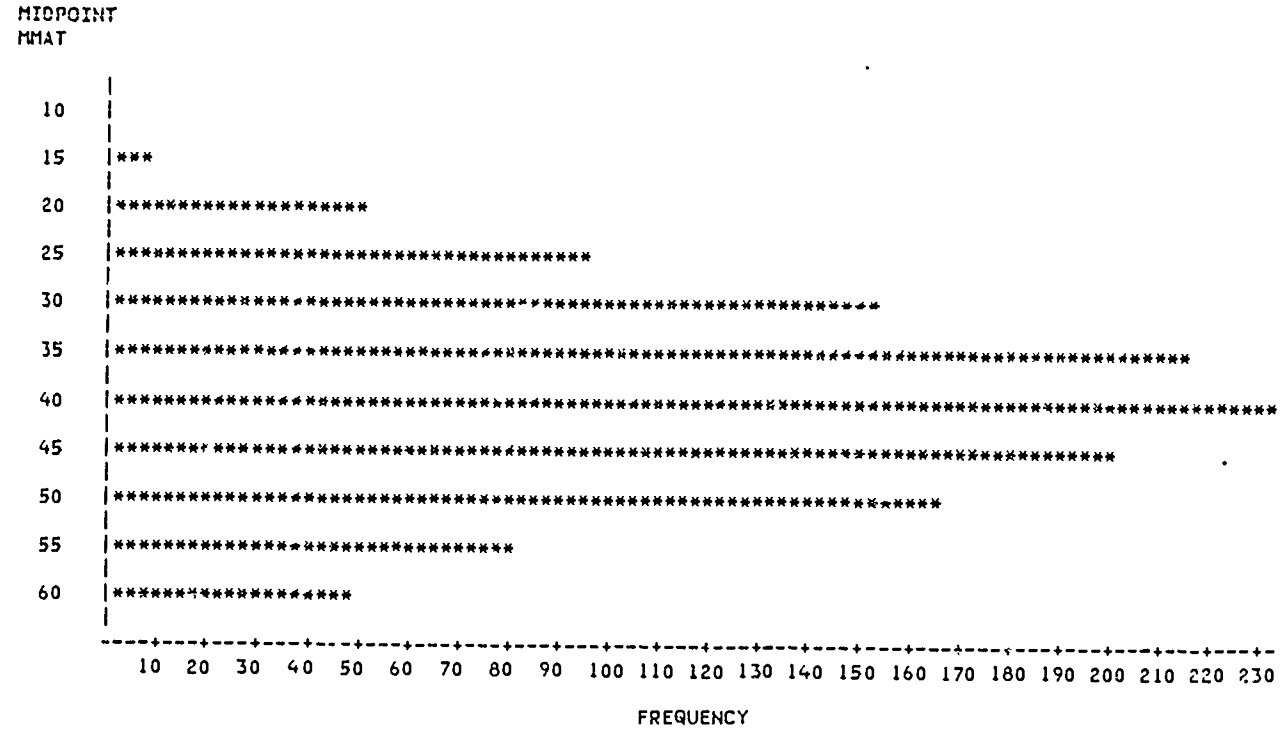


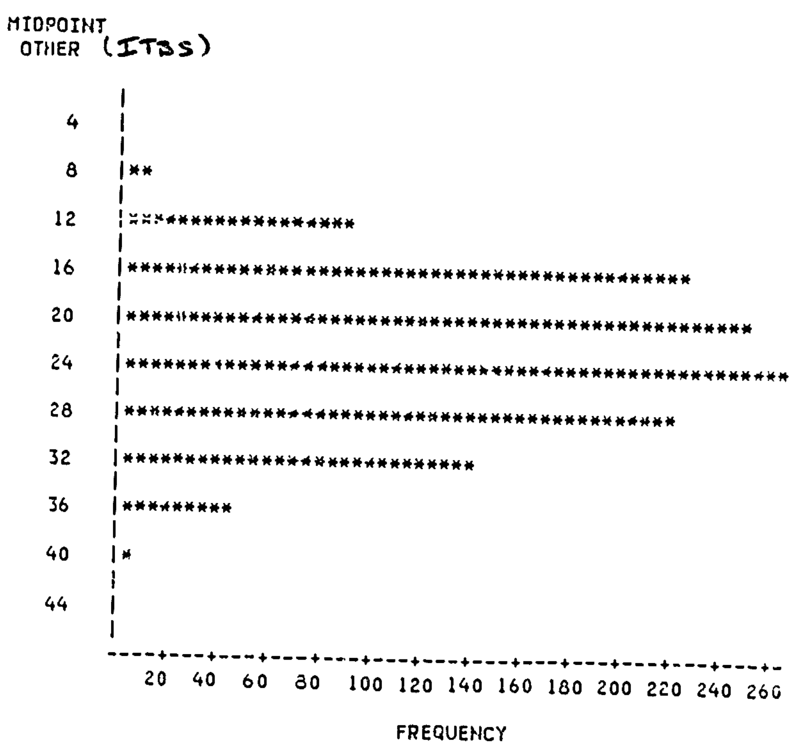
Grade 6 Science

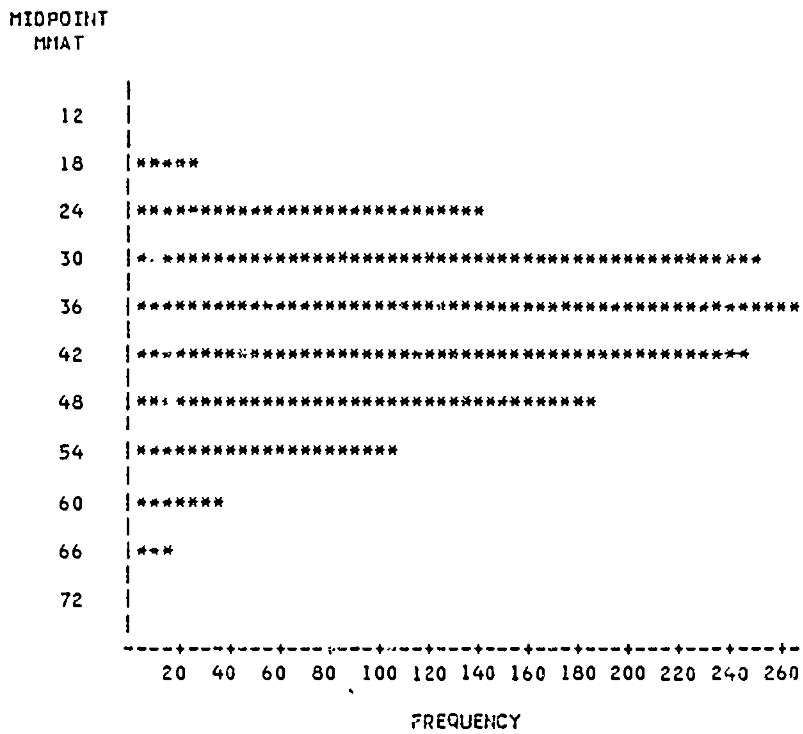


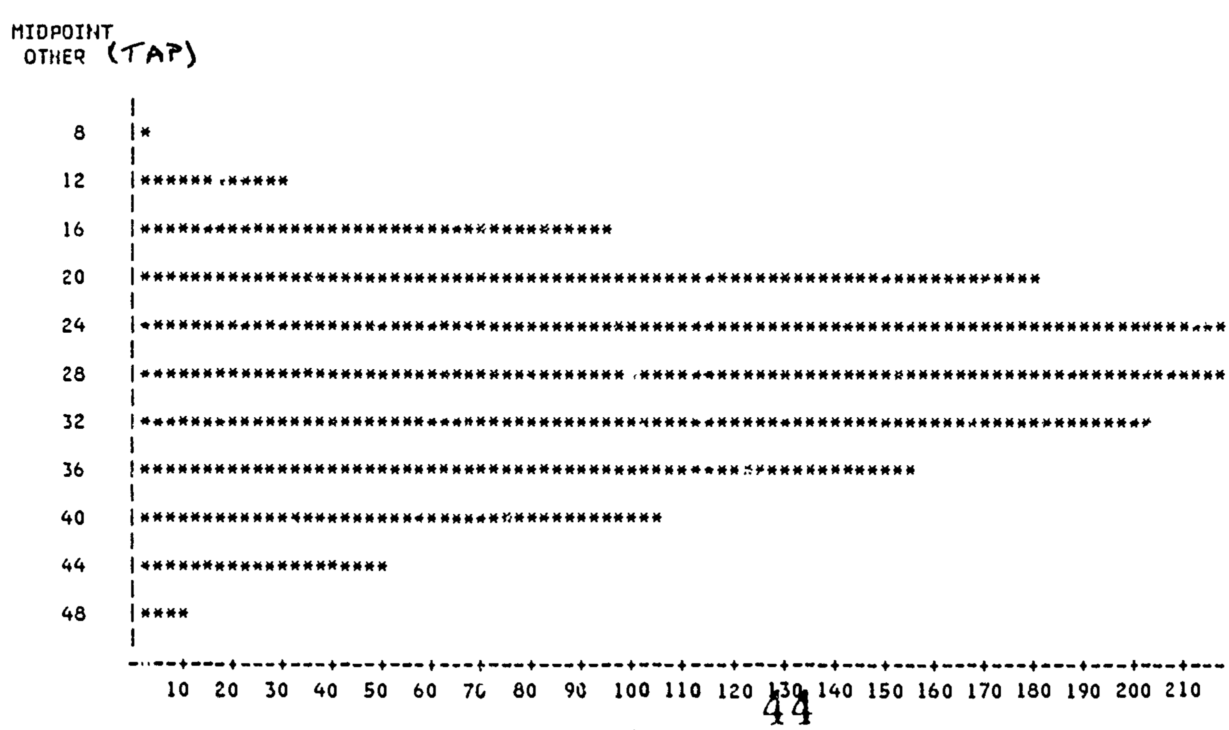


Grade 8 Science

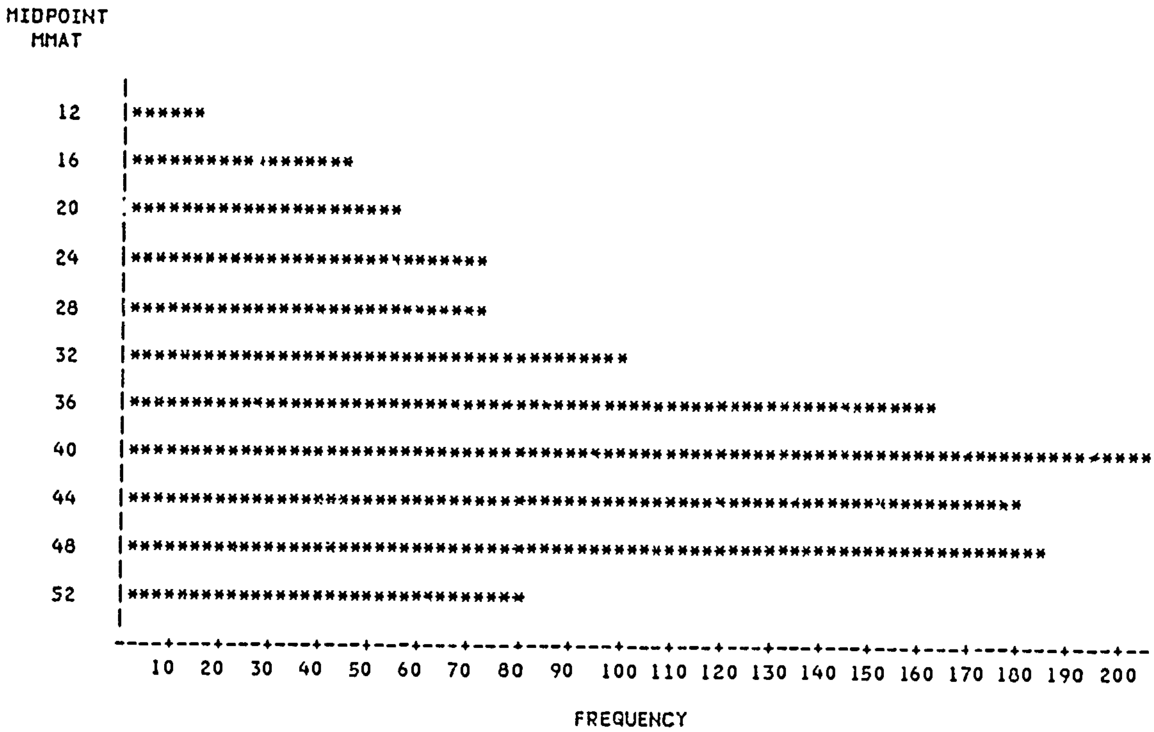


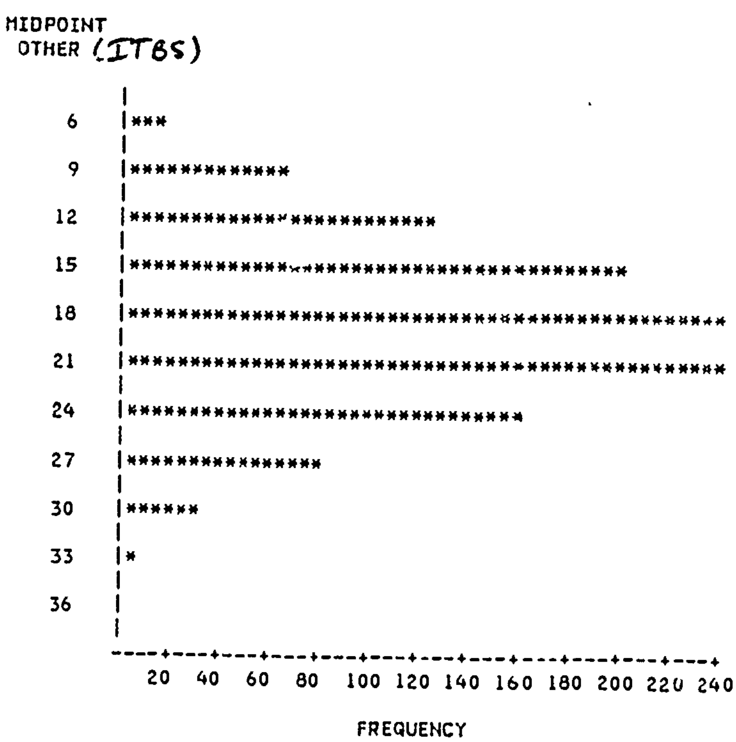



MIDPOINT
MMAT
12
18
24
30
36
42
48
54
60
66
72
20 40 60 80 100 120 140 160 180 200 220 240 260
FREQUENCY

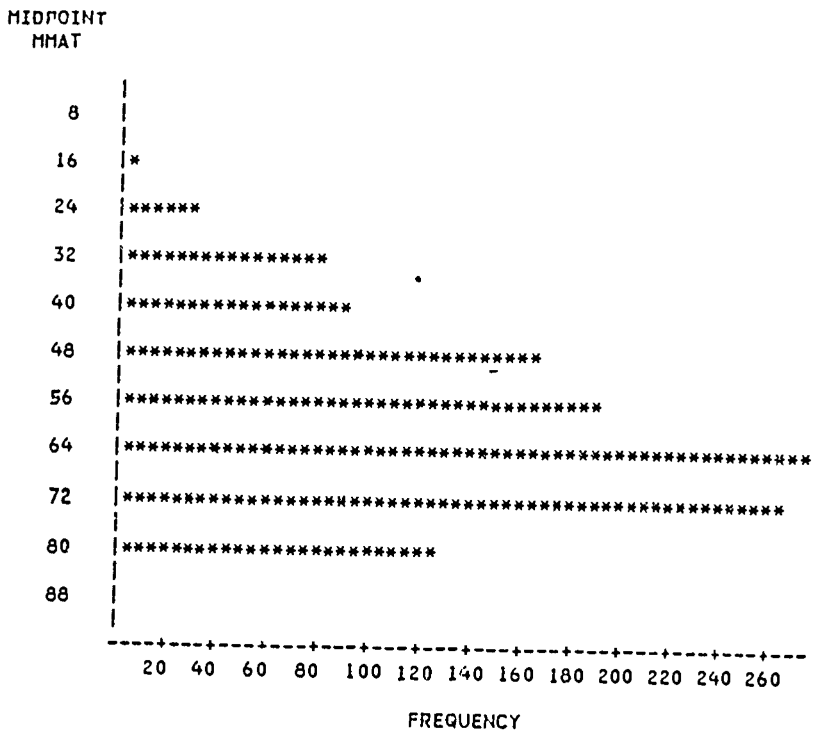

MIDPOINT
OTHER (TAP)
8
12
16
20
24
28
32
36
40
44
48
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210
FREQUENCY

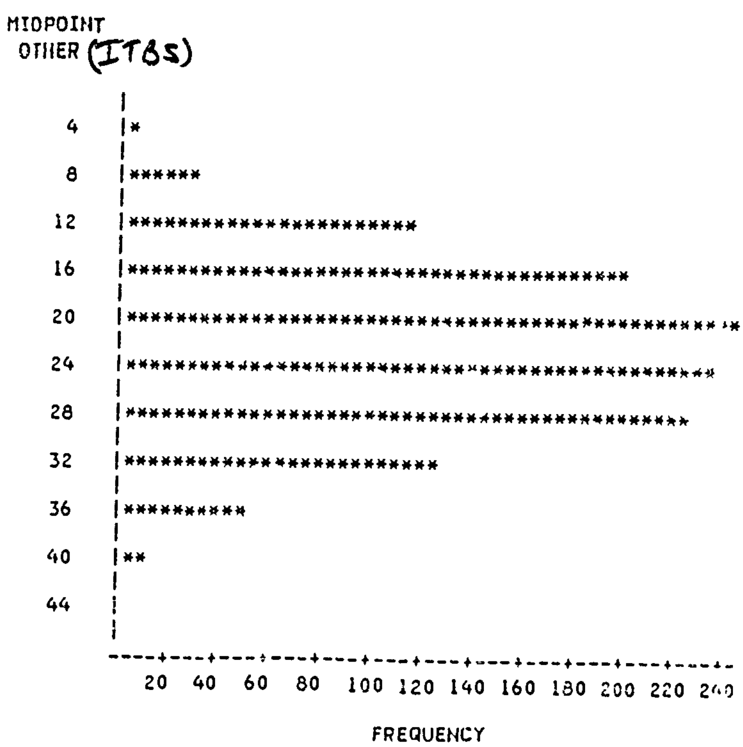
Grade 3 Social Studies



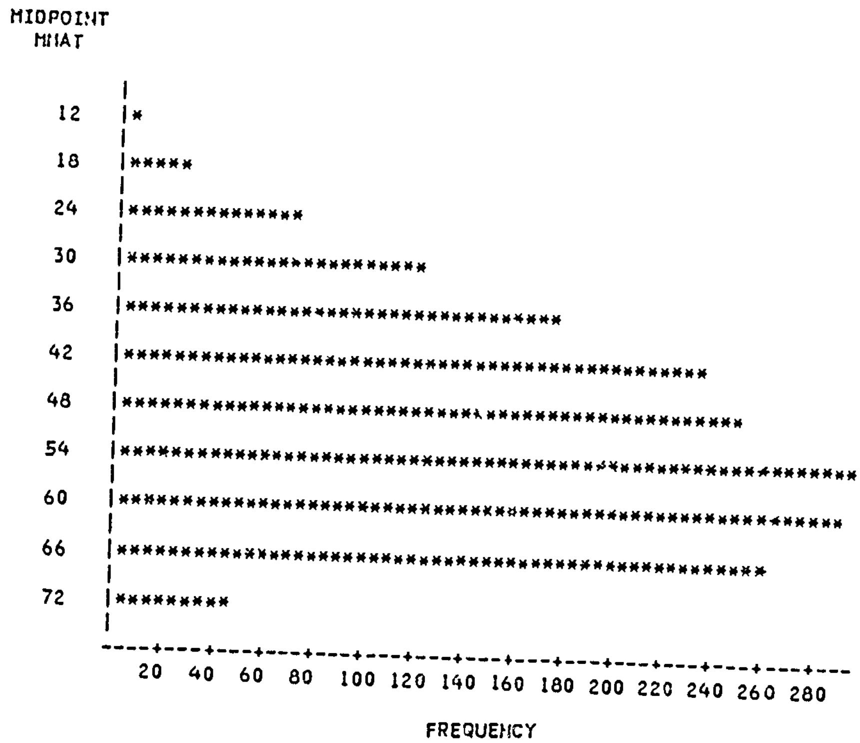


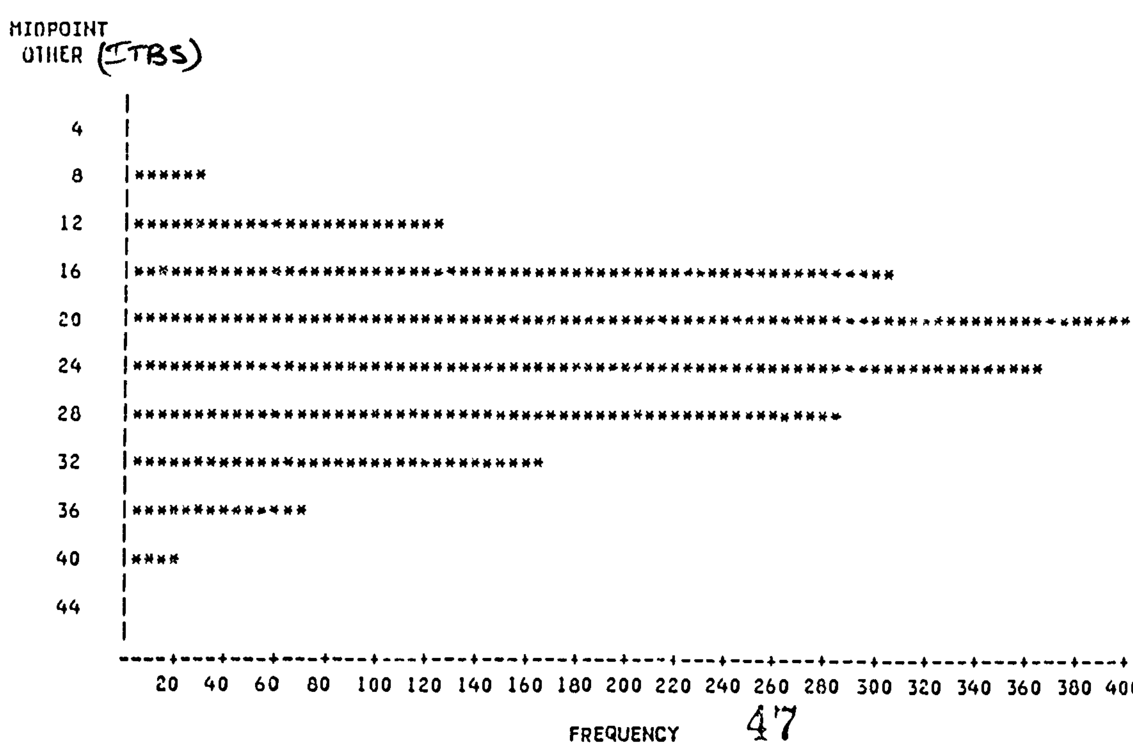
Grade 6 Social Studies



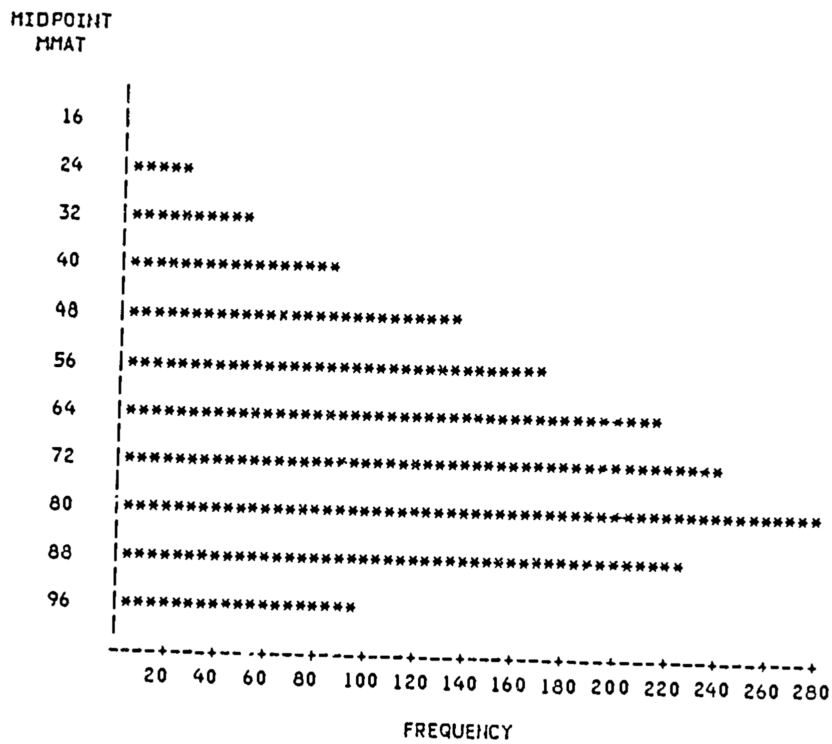


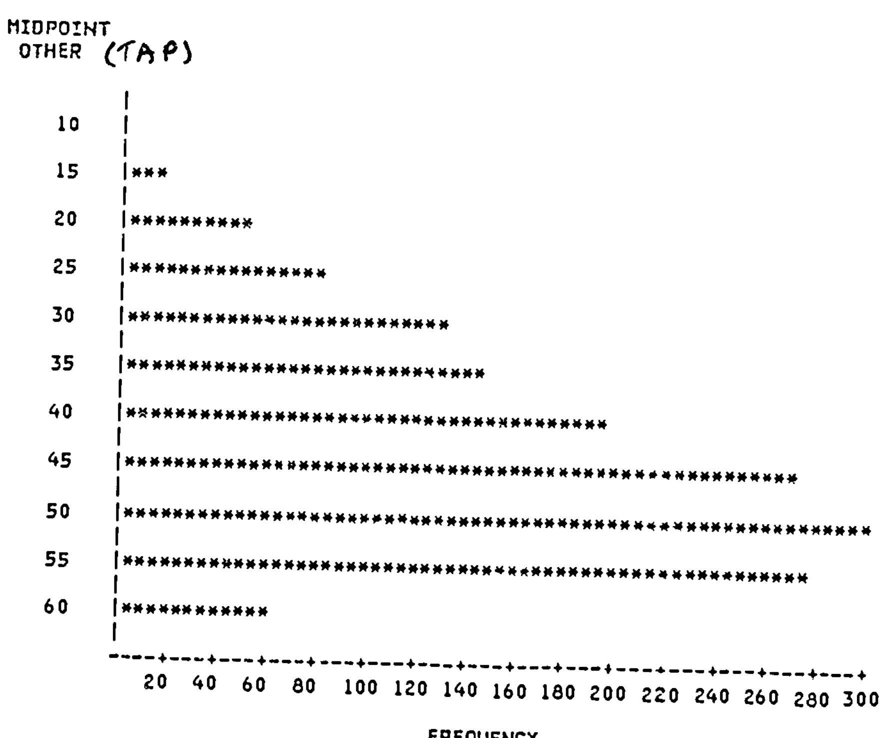
Grade 8 Social Studies





Grade 10 Social Studies



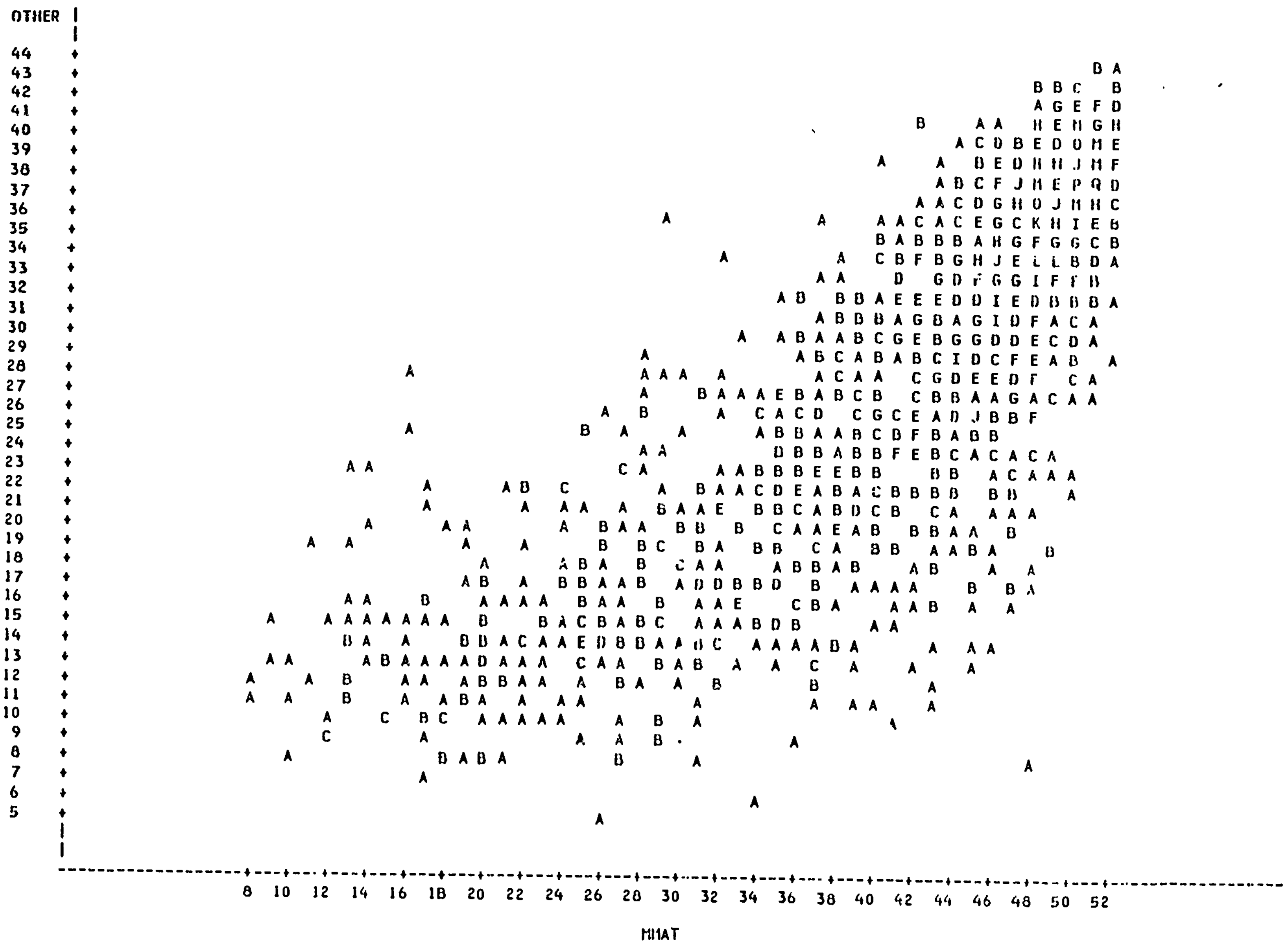


Figures 20 through 38

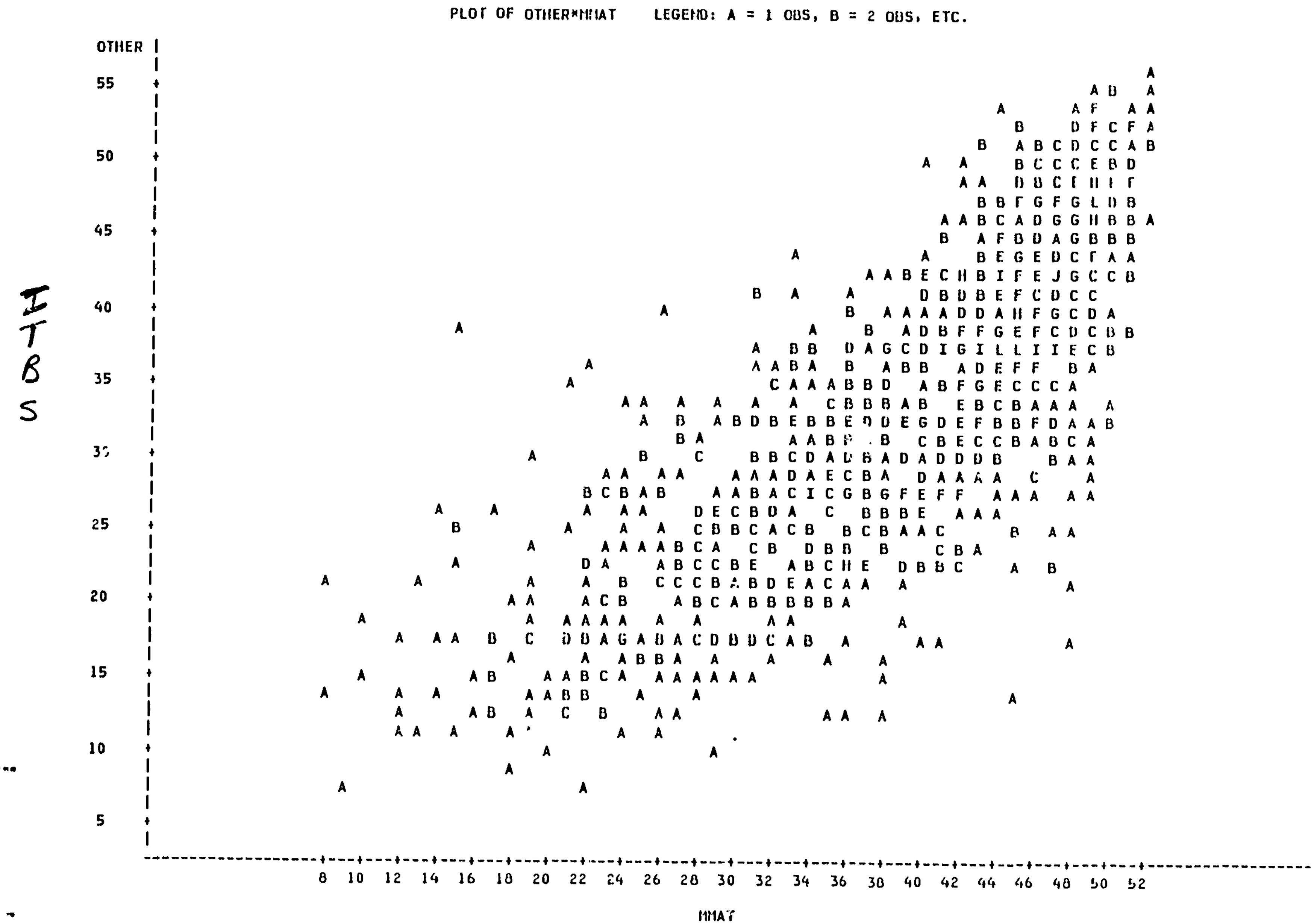
Relationships of Corresponding MMAT and ITBS/TAP Subject Tests

PLOT OF OTHER*MMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.

ITBS

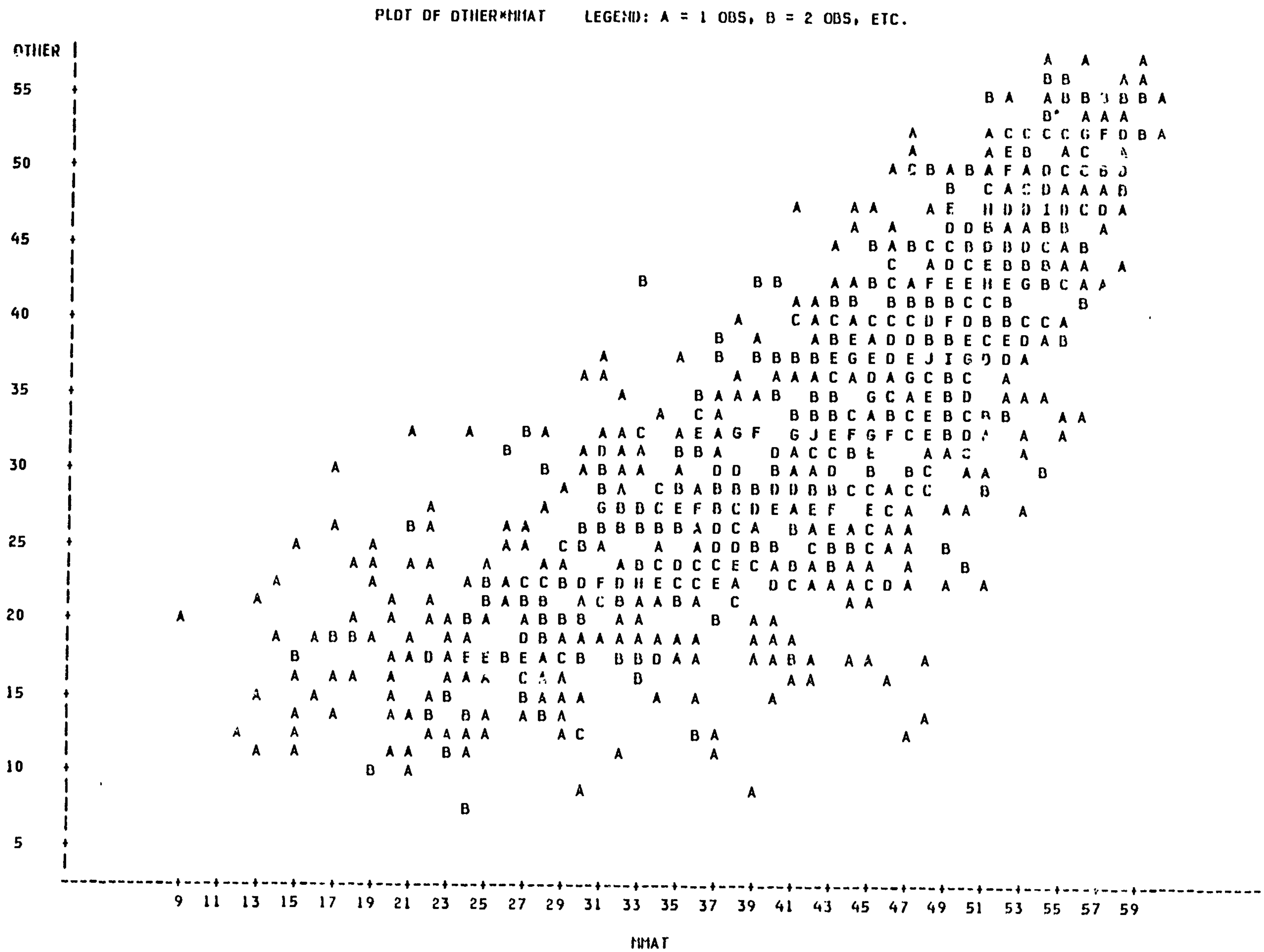


Grade 3 Reading



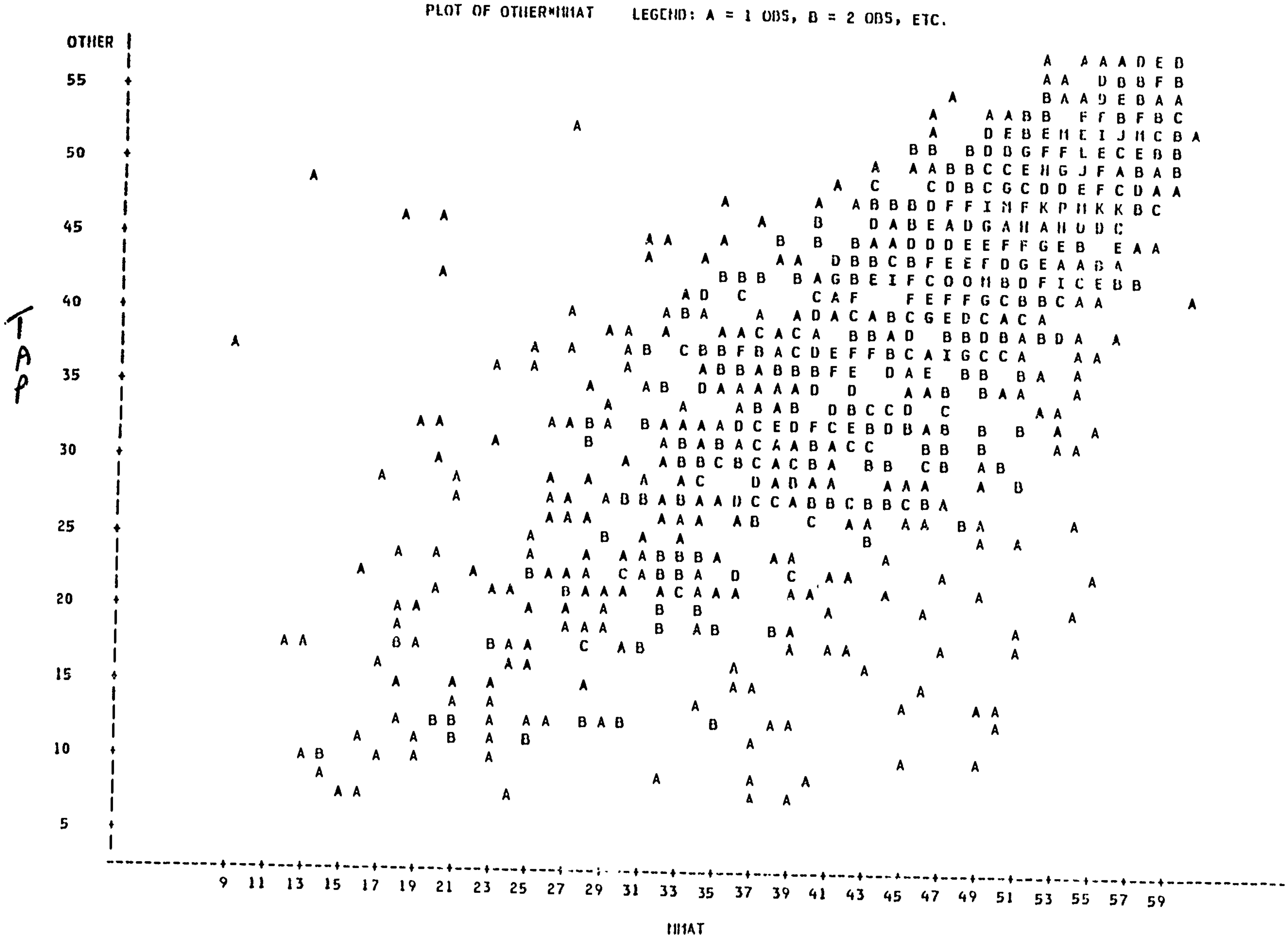
Grade 6 Reading

ITBS

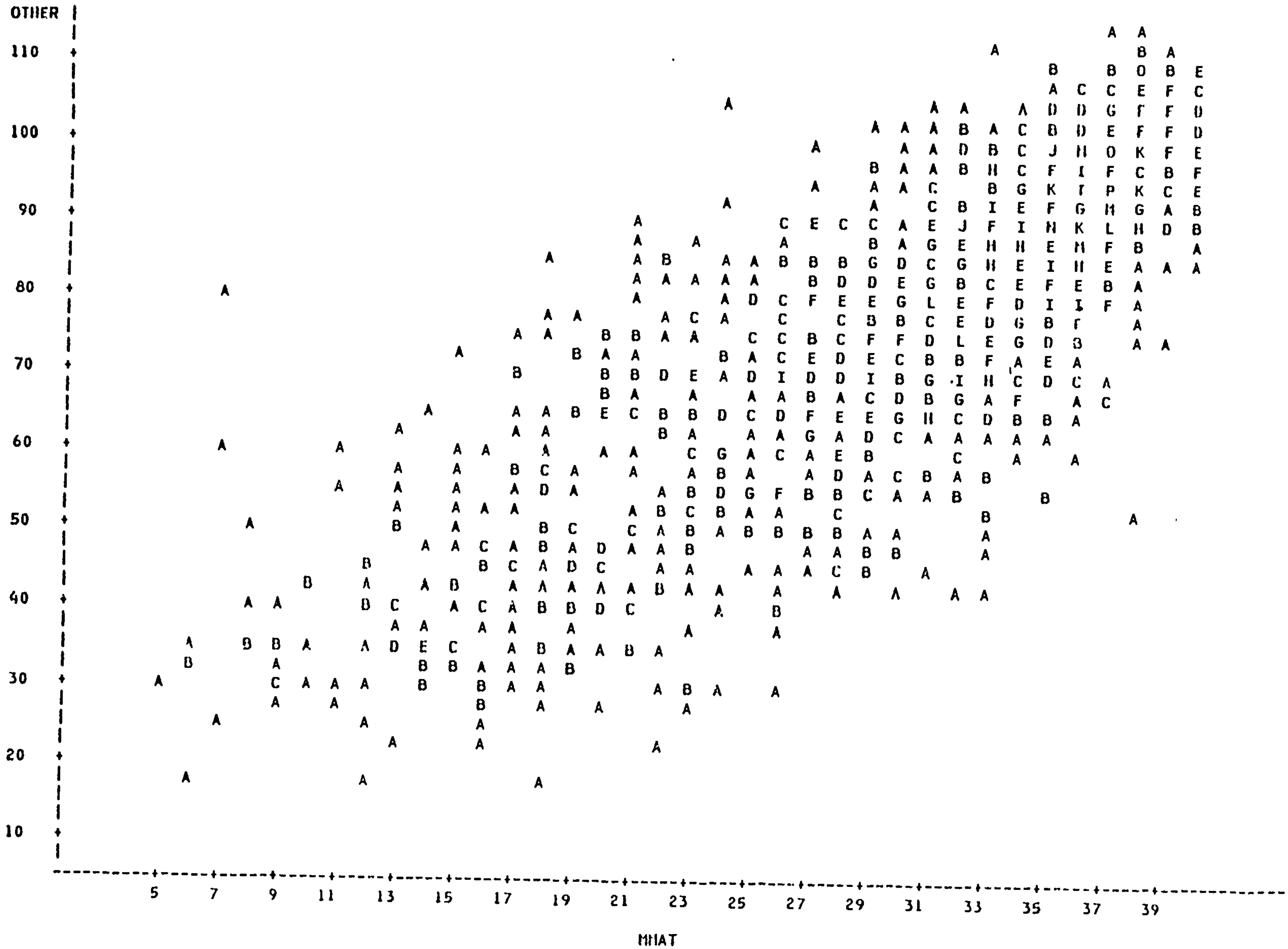


Grade 8 Reading

Grade 10 Reading



PLOT OF OTHER*MNAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



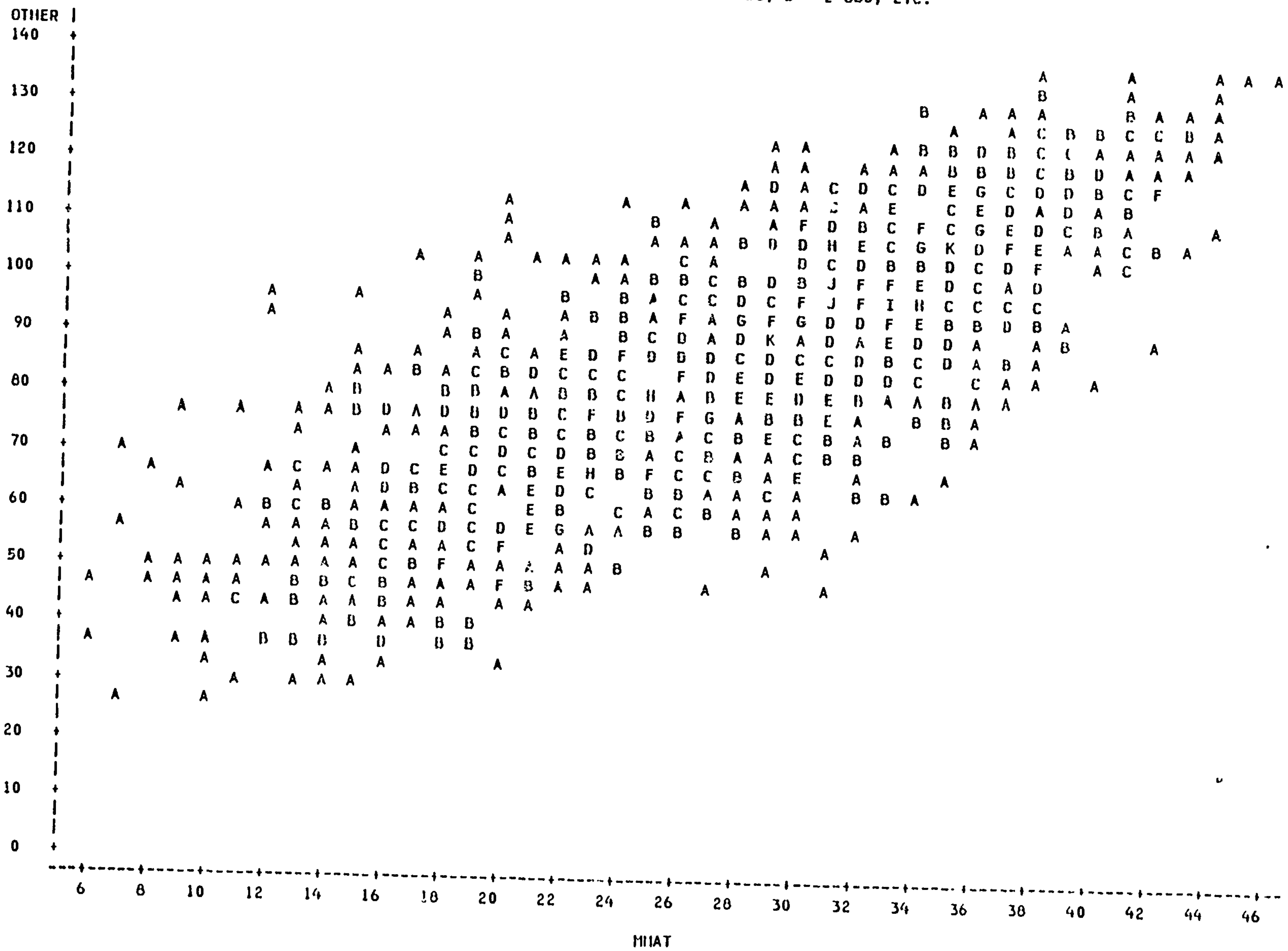
ITBS

Grade 3 Language Arts

52

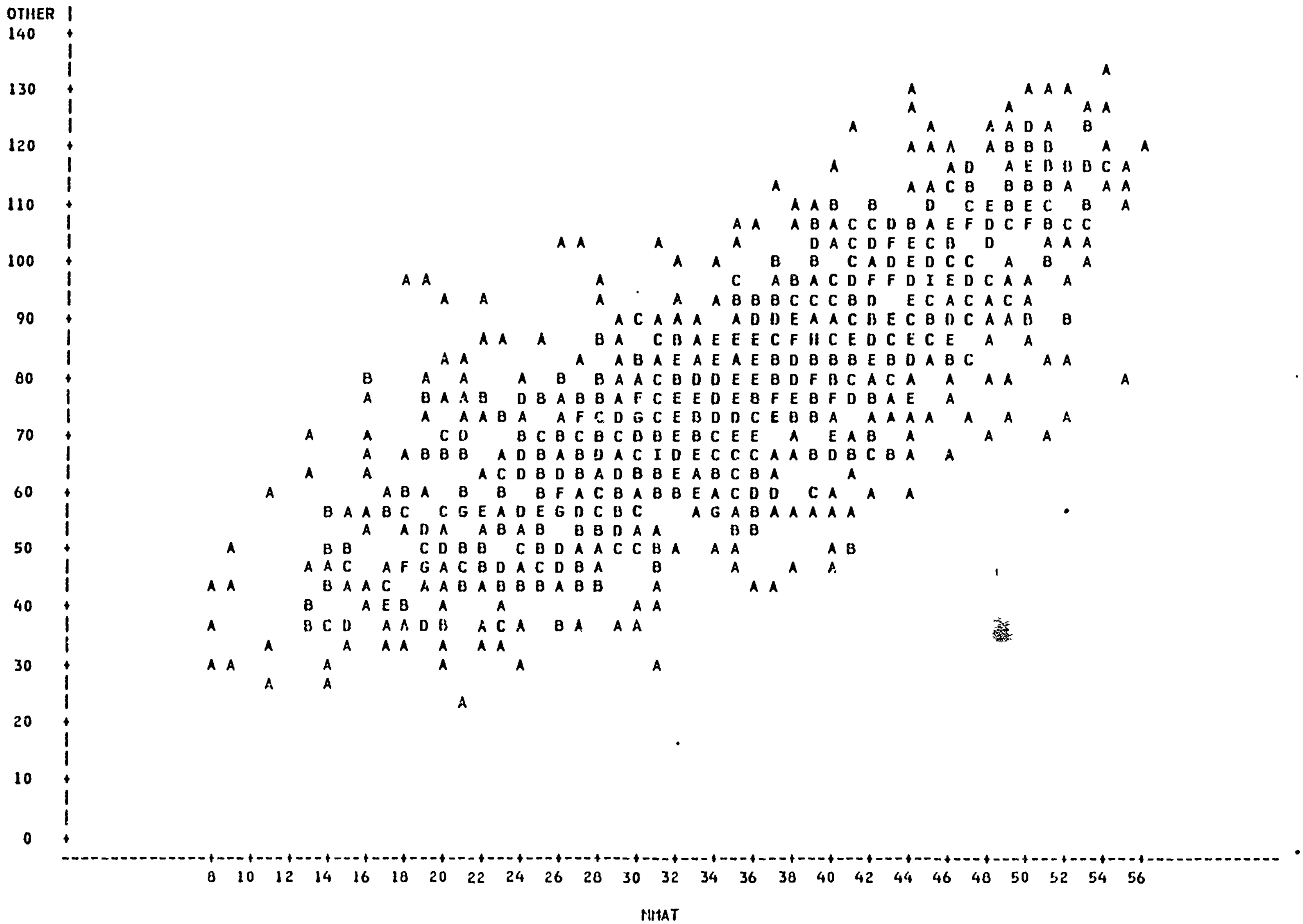
PLOT OF OTHER*MMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.

ITBS



Grade 6 Language Arts

PLOT OF OTHER*NMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.

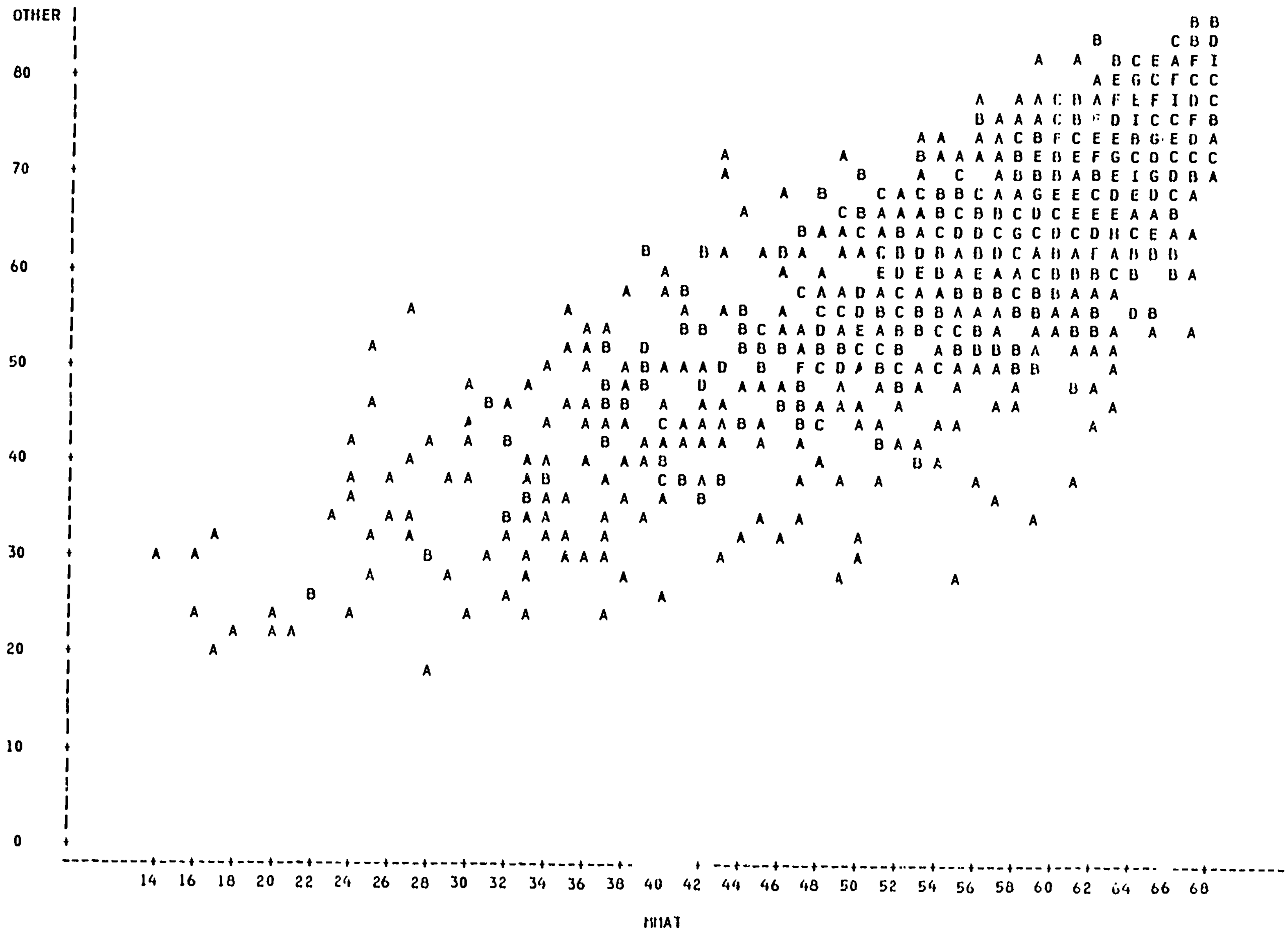


ITBS

Grade 8 Language Arts

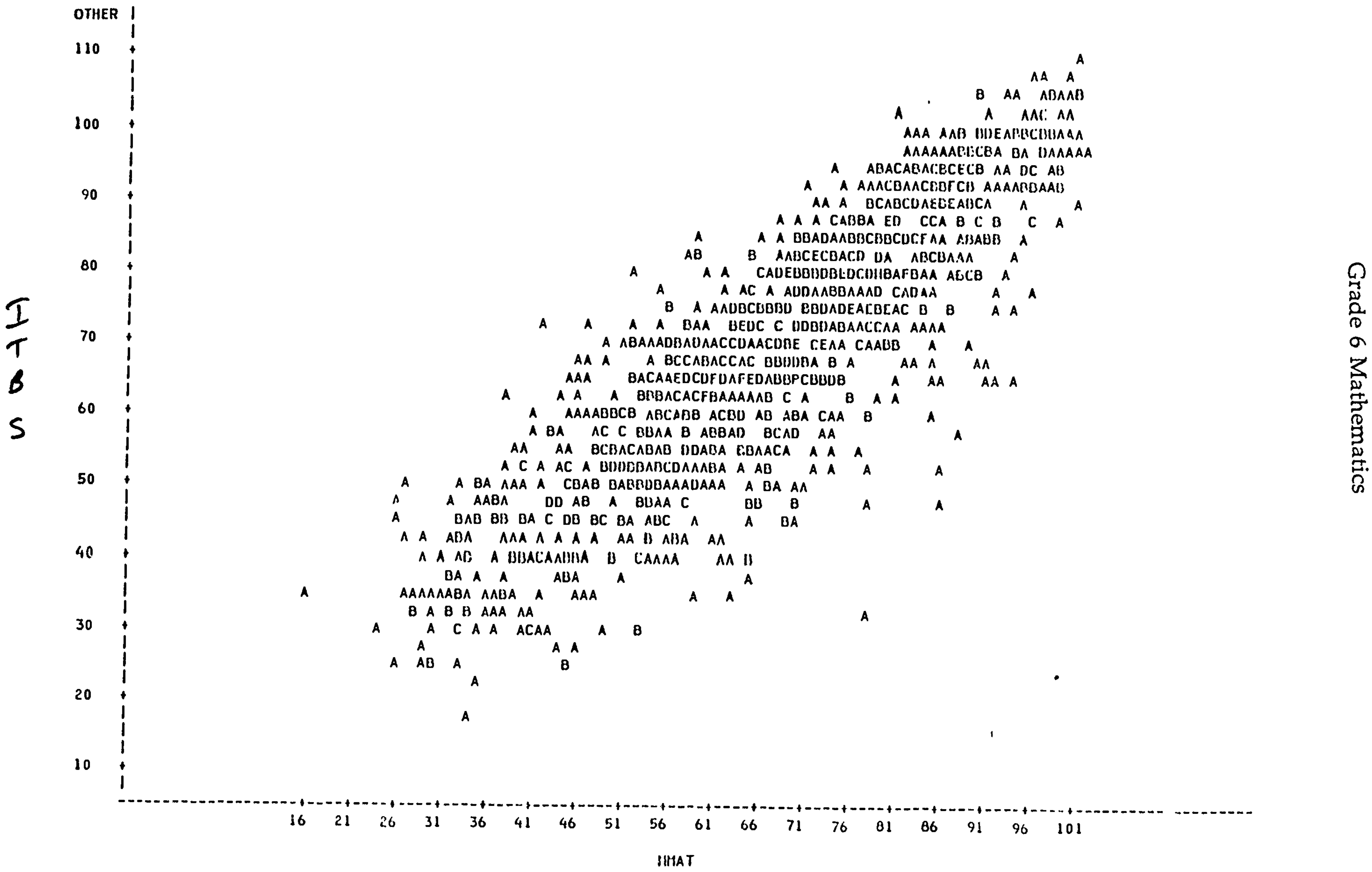
PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.

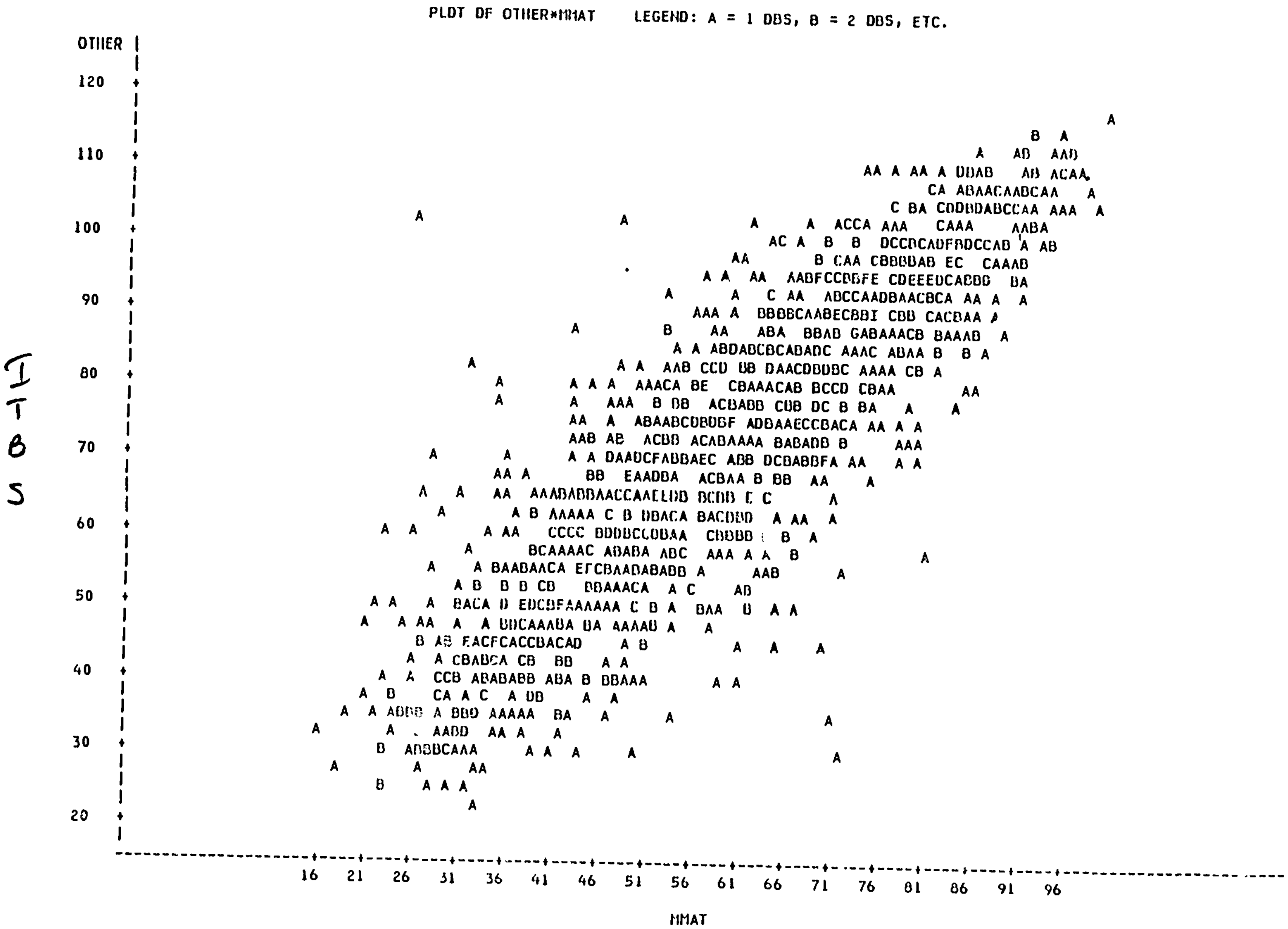
ITBS



Grade 3 Mathematics

PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.

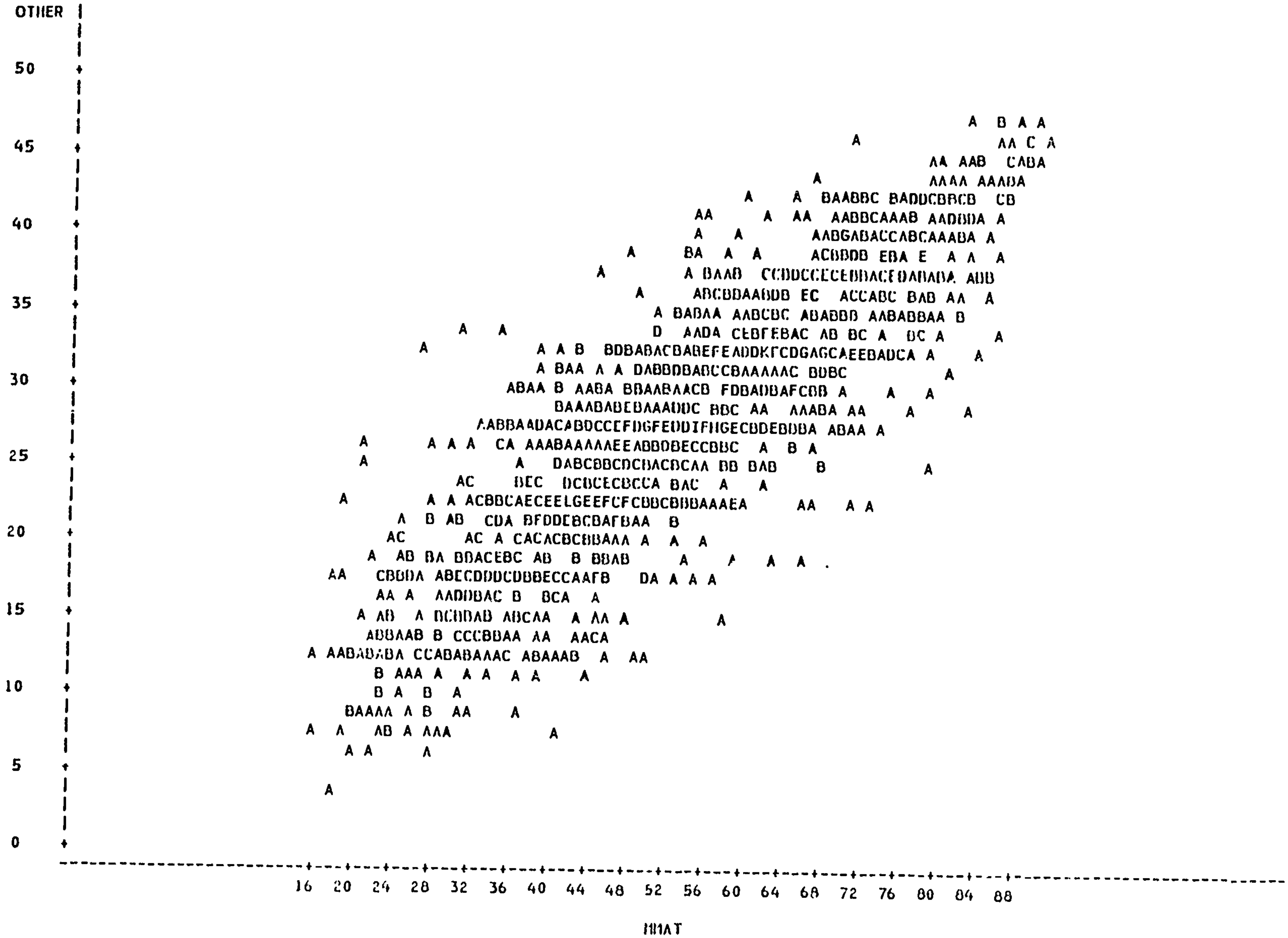




Grade 8 Mathematics

57

PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.

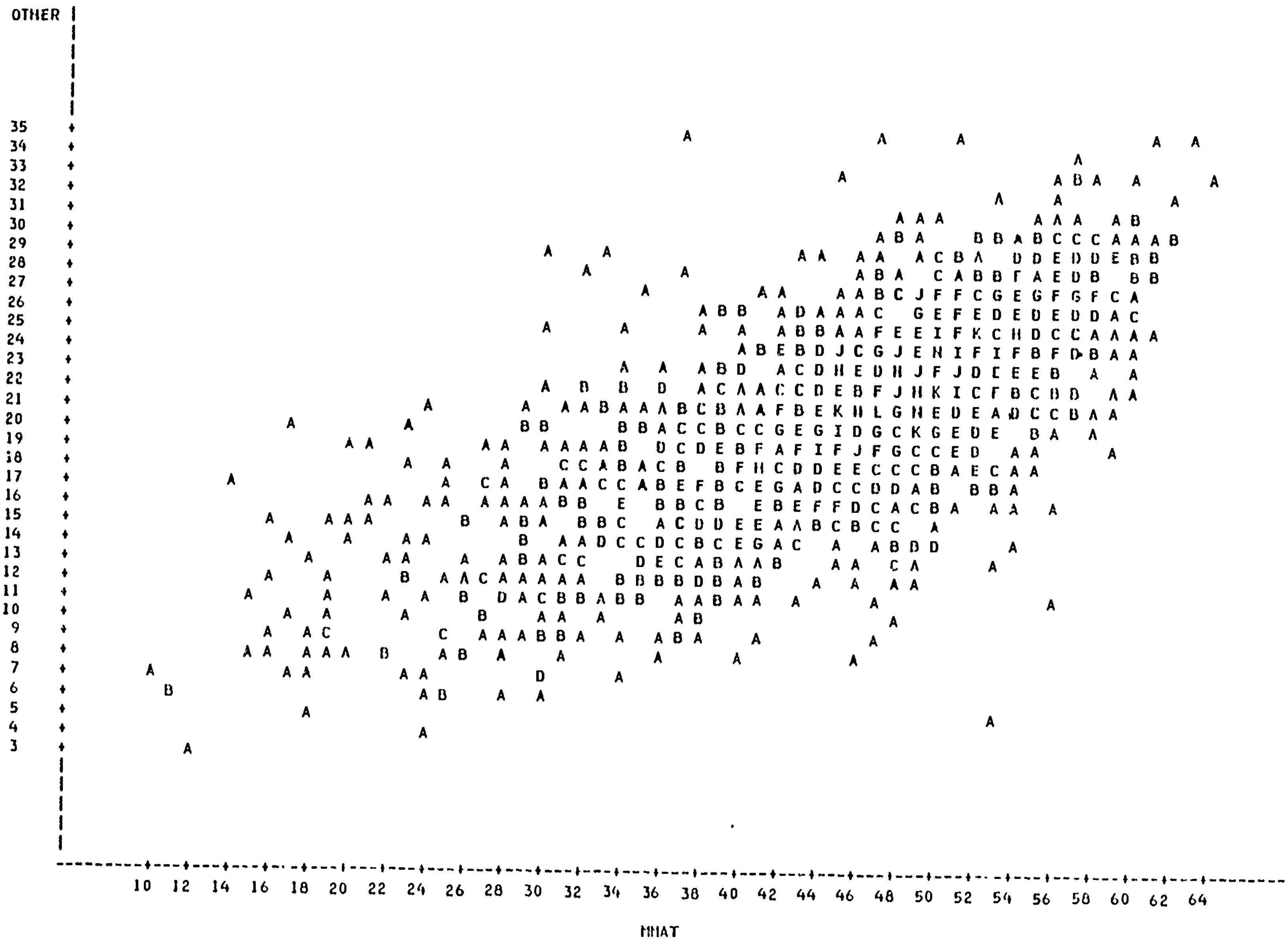


TAP

Grade 10 Mathematics

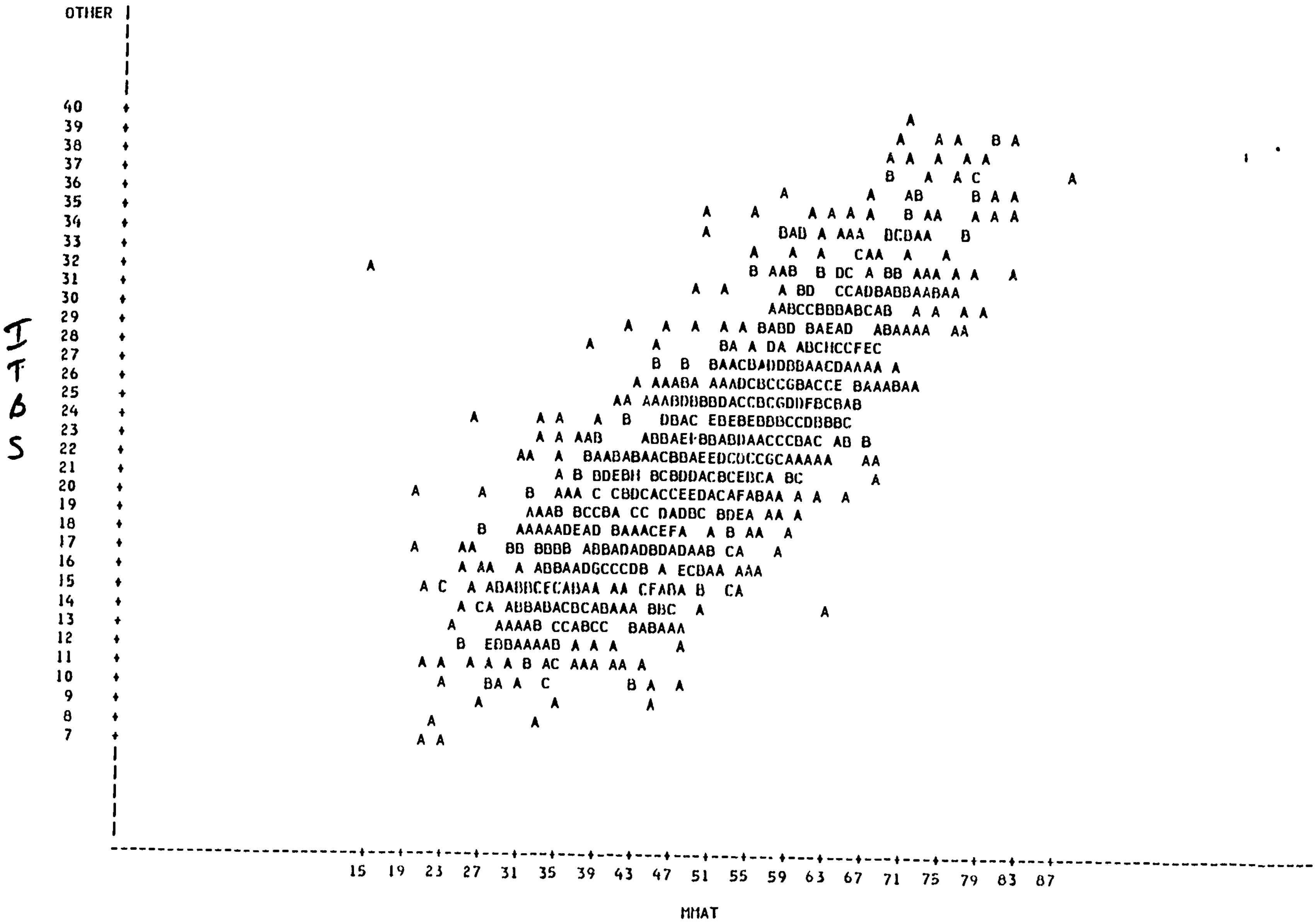
PLOT OF OTHER*MMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.

ITBS



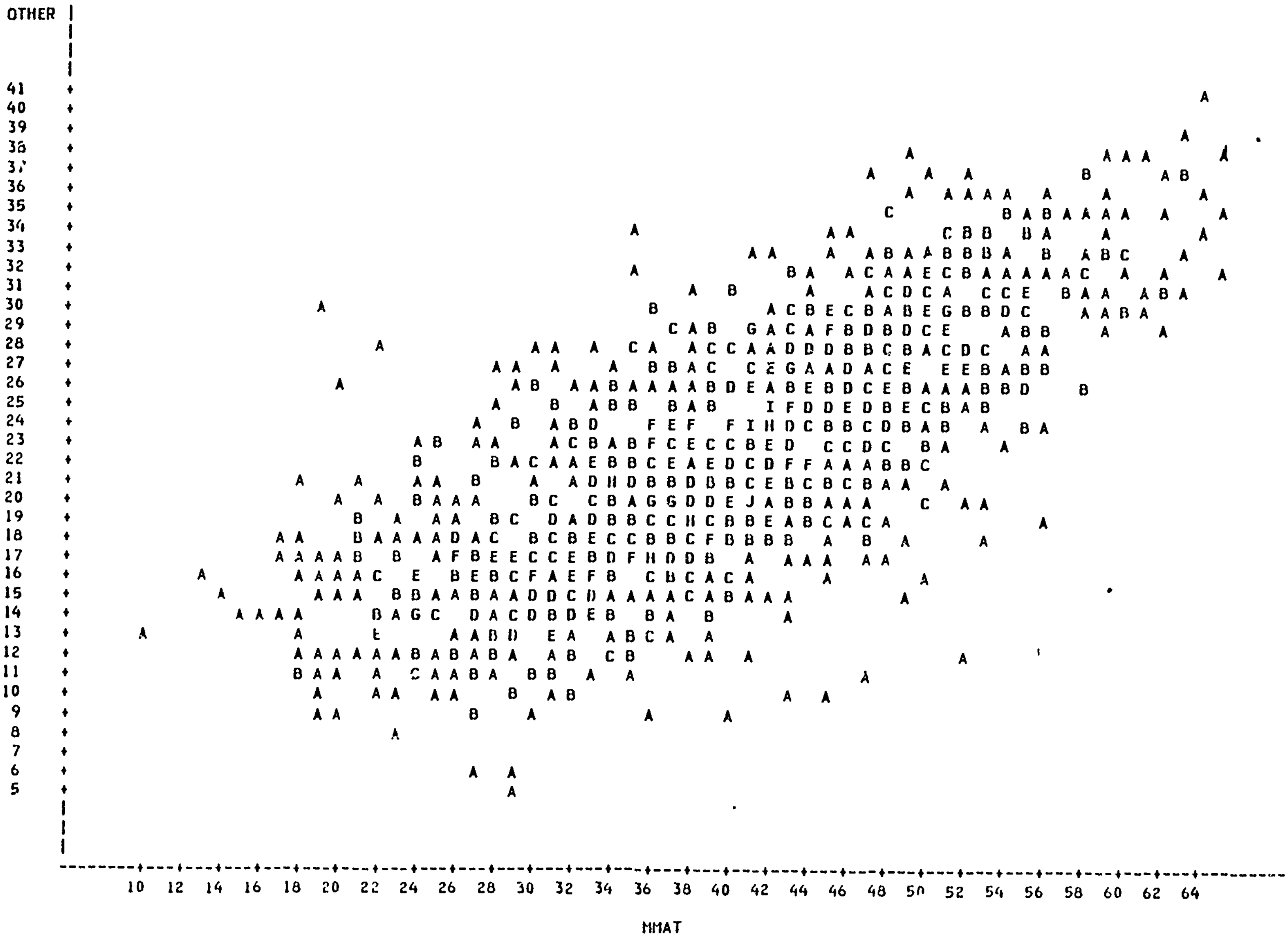
Grade 3 Science

PLOT OF OTHER*MMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



Grade 6 Science

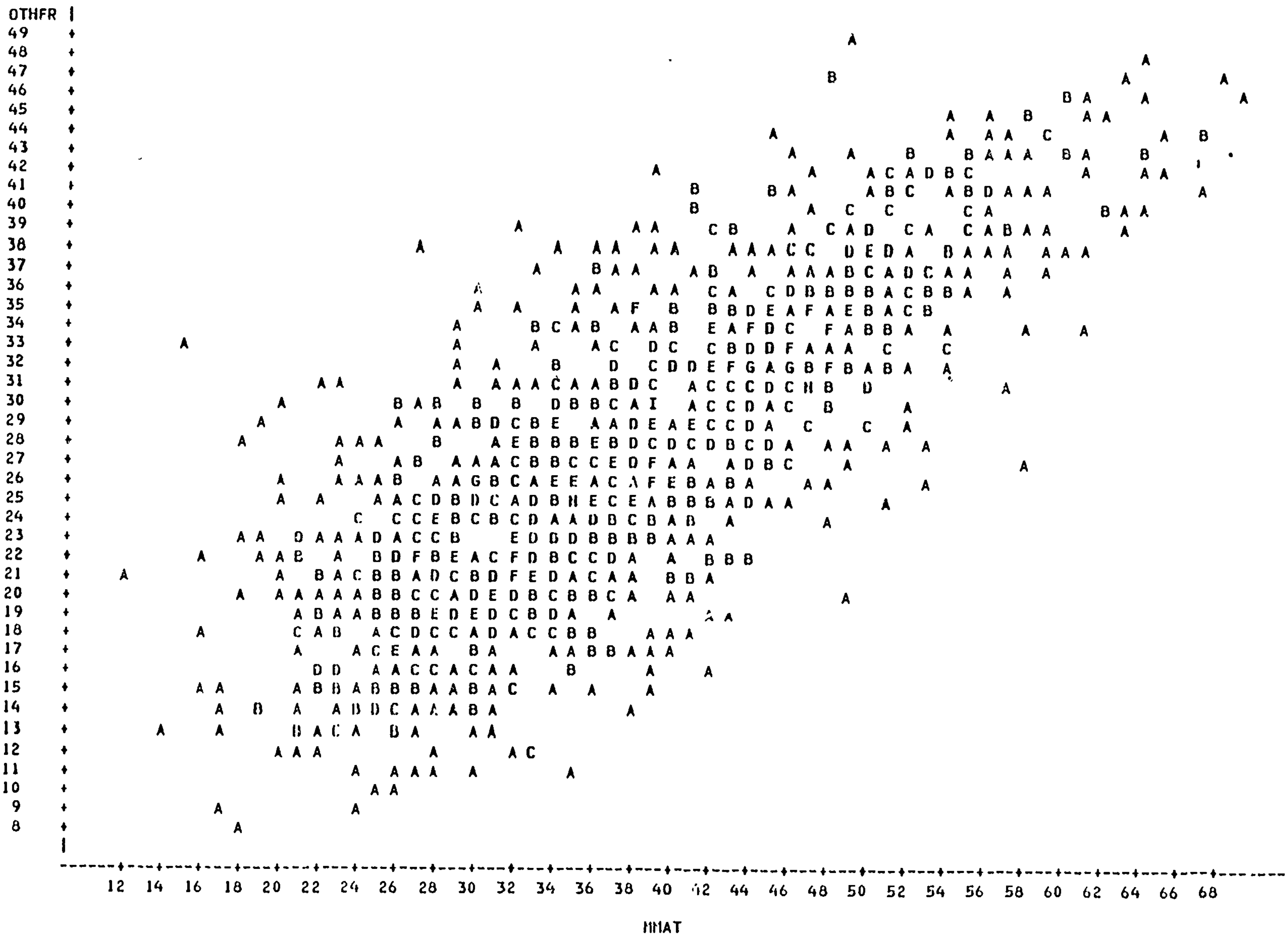
PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.



T/T B S

Grade 8 Science

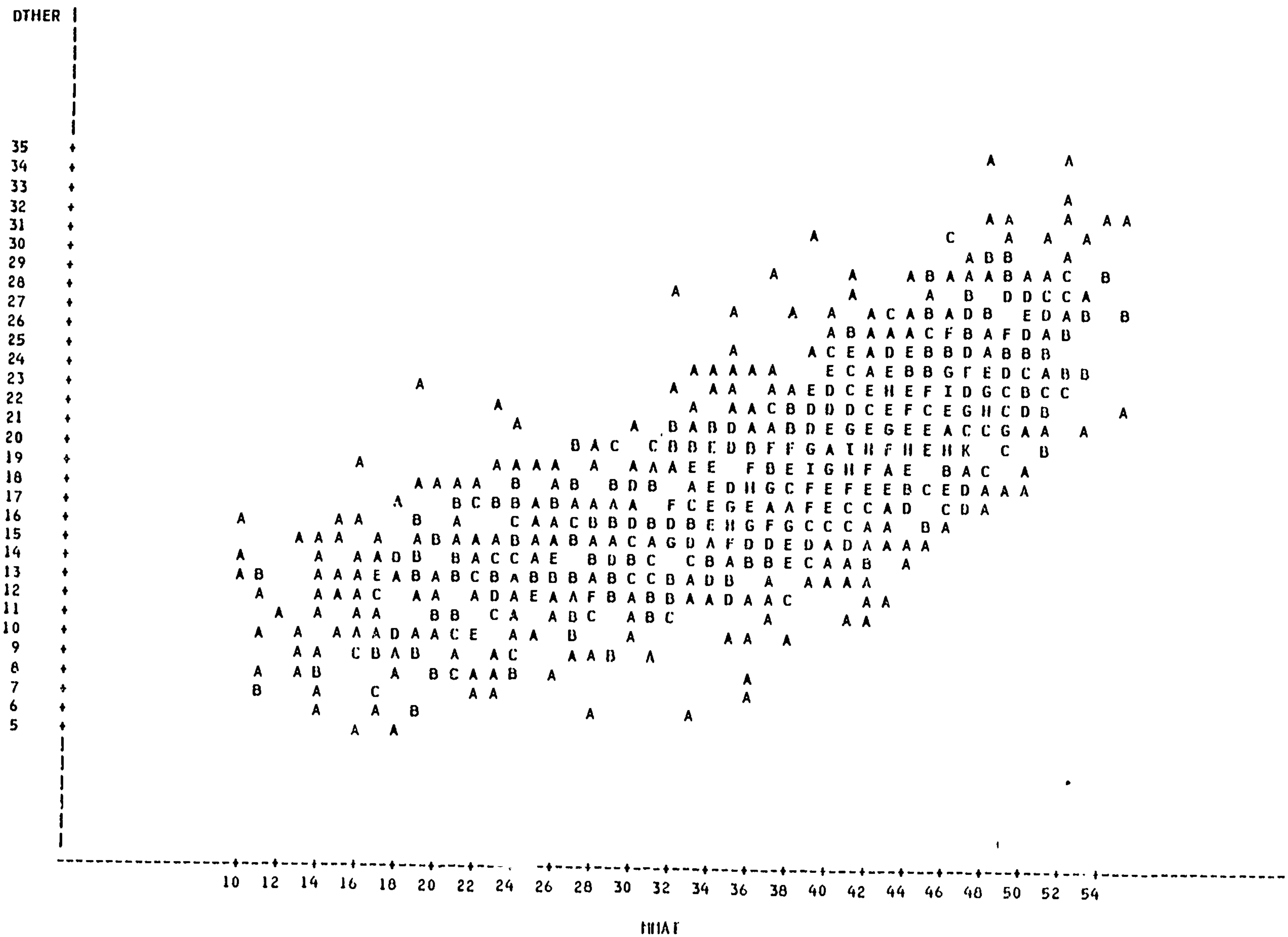
PLOT OF OTHER*MMAT   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



TAP

Grade 10 Science

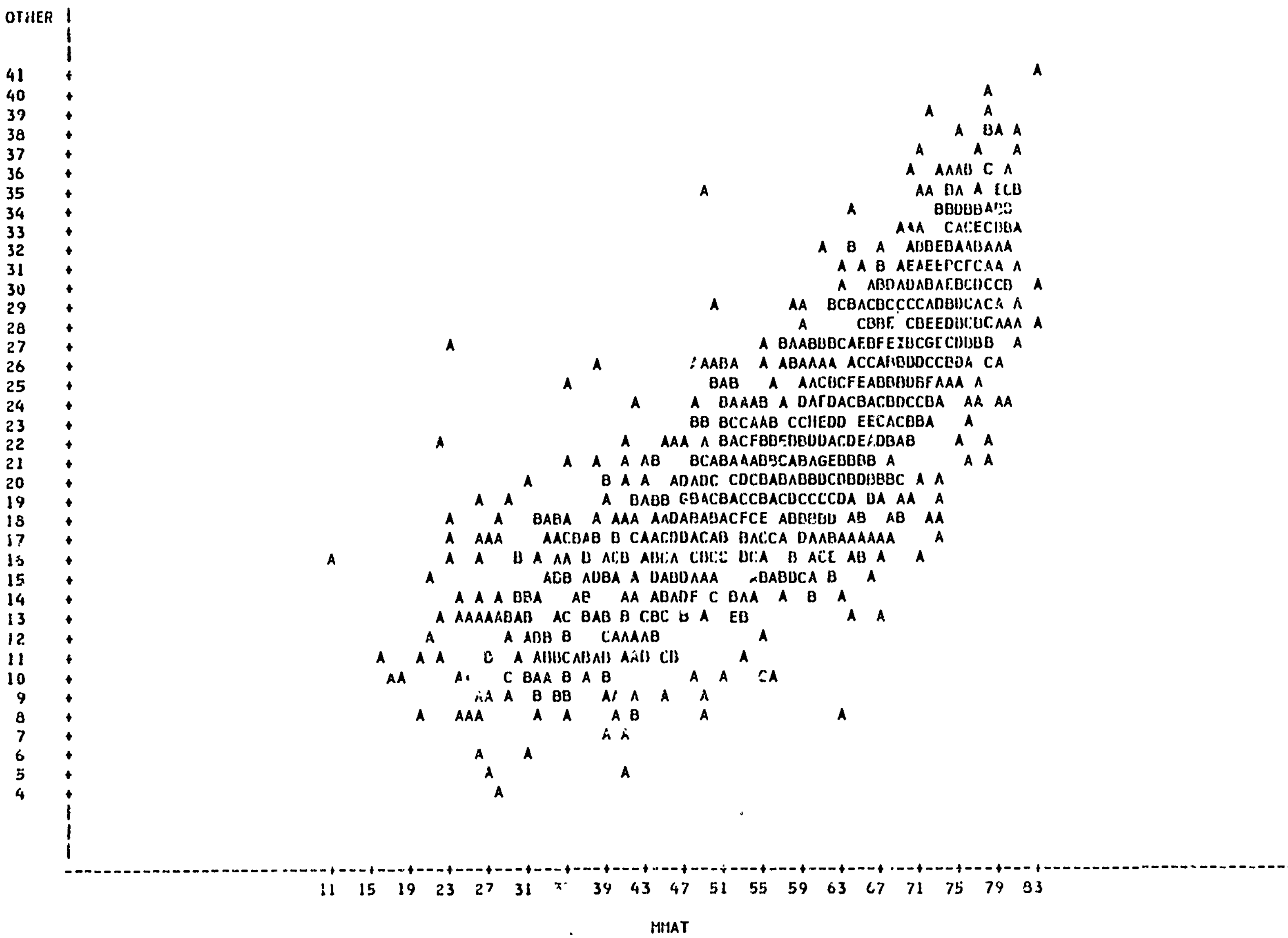
PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.



TTBS

Grade 3 Social Studies

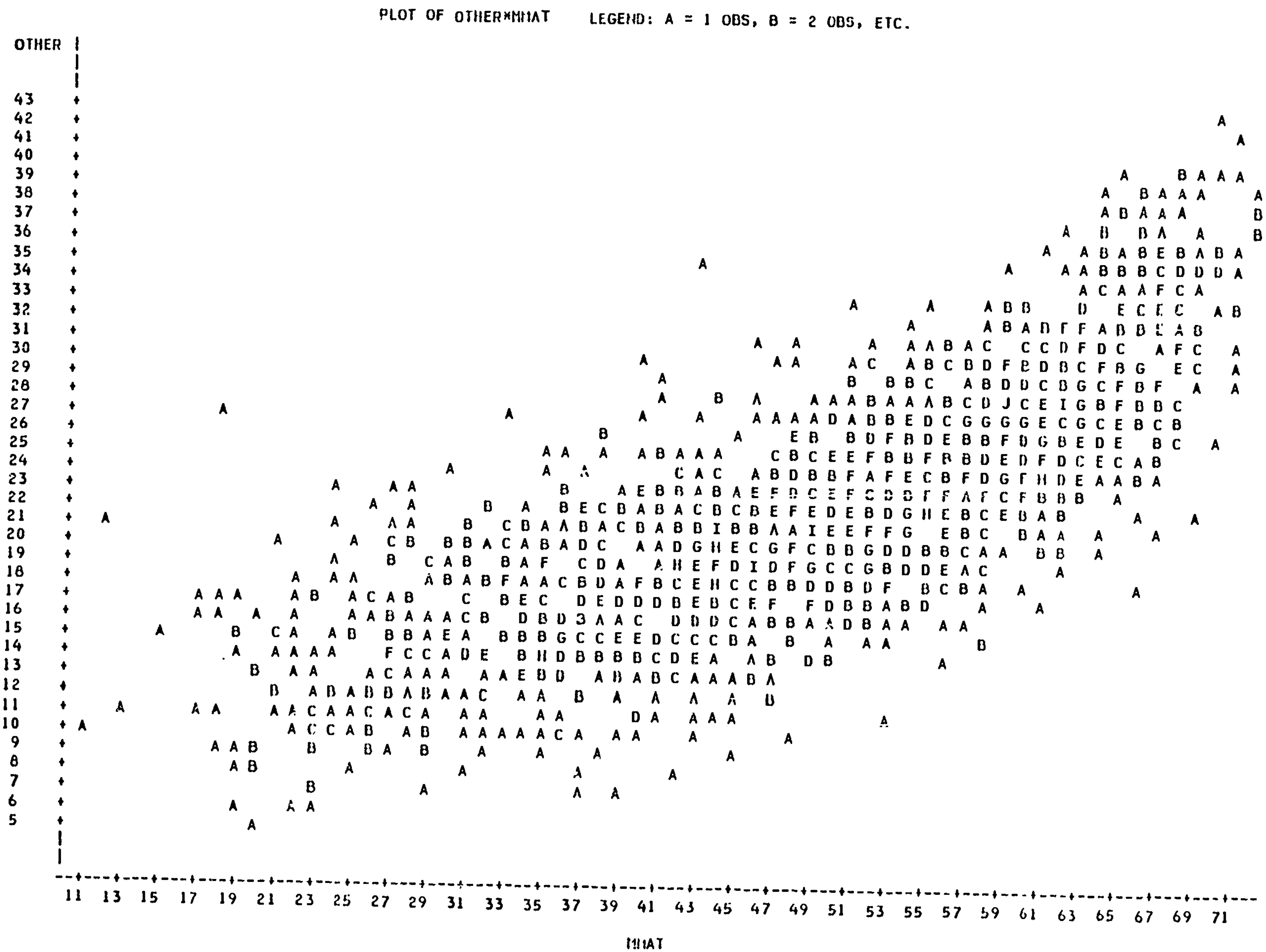
PLOT OF OTHER*MMAT LEGEND: A = 1 OBS, B = 2 OBS, ETC.



6/TBS

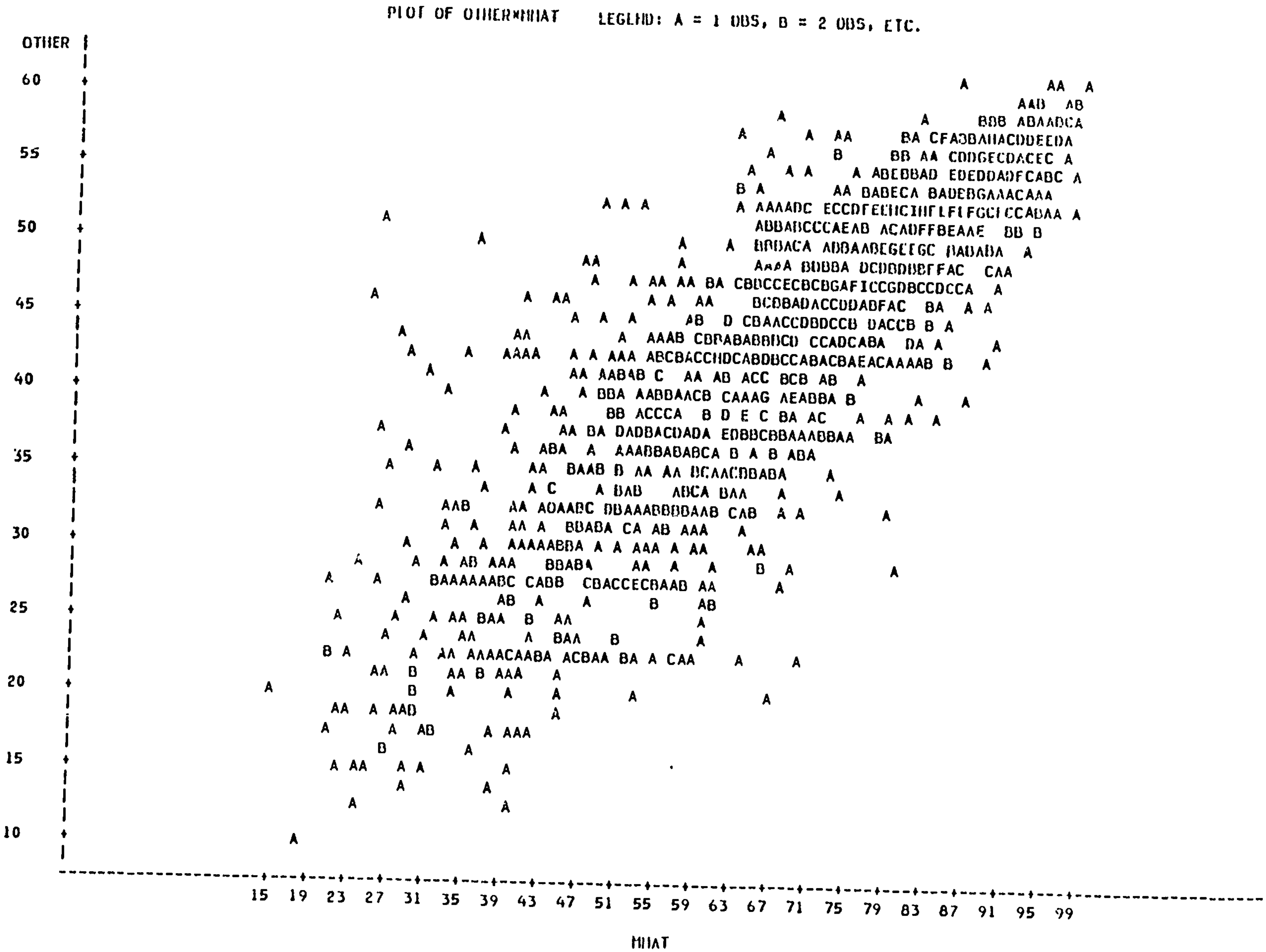
Grade 6 Social Studies

ITBS



Grade 8 Social Studies

TAP



Grade 10 Social Studies

8.6

8.7

Appendix A

Missouri Mastery and Achievement Tests

Technical Summary

The *Missouri Mastery and Achievement Tests* (MMAT) consist of 34 separate objective-referenced tests designed to assess student performance in grades 2 through 10. At all levels but grade 2, these tests measure learner outcomes, called "Key Skills," in four areas: language arts/reading/English, mathematics, science, and social studies/civics. Tests for grade 2 are limited to language arts/reading and mathematics.

Multiple choice items, each with four options, are used on the MMAT. There are three equivalent forms (A, B, and C) at grades 3, 6, 8, and 10 and two equivalent forms (D and E) at grades 2, 4, 5, 7, and 9.

Test Development

Every effort was made during the development of the MMAT to ensure that the battery would meet the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985). At each step, the guiding principle was that the MMAT must yield reliable and valid measures of student achievement.

Experts in each subject area prepared test content specifications in order to provide a set of parameters for measurement of each Key Skill. These specifications were reviewed by elementary, secondary, and college-level teachers. Then a third group of educators used the specifications as blueprints for item writing. Twice as many items were written as were needed for the final forms.

Items were subjected to a thorough editing and review process prior to and concurrently with field testing. After several rounds of editing, each item was reviewed by at least four content experts for congruence (consistency) with its respective Key Skill. At the same time, representatives from several groups reviewed items for ethnic, gender, and cultural bias or stereotyping. Items were field tested on representative samples of Missouri students; approximately 400 students responded to each MMAT item.

The item analysis data yielded by the field trials and the results of the bias and content reviews were used to select items for the final test forms. To be selected for inclusion an item must have been judged to be congruent and free of any content that resulted in stereotyping or bias, as well as have demonstrated acceptable statistical properties.

A common-items equating design was used to ensure that the final forms constructed for each grade level were indeed parallel. Forms A, B, and C for grades 3, 6, 8, and 10 were subjected to a field trial during the fall of 1986 in order to obtain information about testing time, to try out administration procedures, and to verify equivalency of forms within a grade level.

1987 Administration of Form A

Form A was administered in the Spring of 1987 to approximately 240,000 students in grades 3, 6, 8 and 10. At each of those four grade levels, about 6,000 students were designated to be part of the "state sample"--a representative group of Missouri students. A stratified random cluster

technique was used to select students for participation in the sample. Their scores were used to report performance on the Key Skills to the Missouri General Assembly and for various technical analyses.

Scales

Three types of scores are reported for individual students taking the MMAT· Key Skill mastery, a scaled score for each subject and cluster (clusters are natural grouping of Key Skills within a subject), and an estimated comparable national percentile rank for each subject.

Key Skill Mastery

Each Key Skill is measured by four items. A student must answer at least three items correctly in order to master the Key Skill.

Subject and Cluster Scaled Scores

Subject and cluster scaled scores are derived using item response theory. This type of scaling yields more accurate results than the commonly used number correct scaling (which is based on classical test theory). The two parameter logistic model, as implemented by BILOG (Mislevy and Bock, 1984), is used to compute subject and cluster scaled scores. A student's scaled score represents the relationship of the student's pattern of responses on the entire set of items to the specific characteristics of each item.

The mean and standard deviation of subject and cluster scaled score distributions are 300 and 65 respectively; the range is from 40 to 560. Subject scaled scores are computed independently of cluster scaled scores. A subject scaled score is not, for example, the average of its cluster scaled scores.

Estimated Comparable National Percentile Ranks

Estimated comparable national percentile ranks are based on the performances of students in the state sample on the MMAT and the *Iowa Tests of Basic Skills* (ITBS) at grades 3, 6, and 8 or the *Tests of Achievement and Proficiency* (TAP) at grade 10. The process used to obtain comparable scores involves relating the sample's distribution of MMAT raw scores to the sample's distribution of ITBS or TAP raw scores. The comparable national percentile ranks derived from state sample data are then used to estimate comparable percentile ranks for all students taking the MMAT.

Each student in grades 3, 6, and 8 receives an estimated comparable national percentile rank in reading/English, language arts, mathematics, science, and social studies. Each tenth grade student receives this type of score for all of the above except language arts.

Score Reports

The MMAT is especially flexible as an educational tool because scores are reported through a variety of complementary forms, all appropriate to different educational situations. The Individual Student Report presents one student's results for each subject and its associated clusters as well as Key Skill mastery data. The Pupil List Report is a roster of all students in a grade; it presents Key Skill mastery information for every student. The Grade Level Key Skill Report, the Grade Level Cluster Report, and the Summary Report present aggregate data for Key Skills, clusters, and subjects respectively at the building and district level. The Chapter I Eligibility List presents students eligible in reading, language arts, and mathematics for Chapter I services.

Validity

Content validity was built into the MMAT during the development process because content experts wrote the test content specifications and the test items. In order to further ensure content validity, at least three content experts reviewed each item in order to determine whether it was congruent with its respective Key Skill. All items selected for the final forms were judged to be valid Key Skill measures by at least four content experts.

Evidence for the construct validity of the MMAT is currently being collected through factor analytic and correlational studies. These initial studies are being conducted on the data from the Spring 1987 administration of Form A.

Reliability

Several types of estimates of score reliability were computed on the Form A data, including indices of internal consistency for raw scores, item response theory reliability estimates for subject and cluster scaled scores, and estimates indicating the reliability of Key Skill mastery classifications.

Internal consistency estimates (KR-20) of subject raw score reliability range from .846 to .950 across grades and subjects, with the median estimate equal to .933. Item response theory estimates of cluster scaled score reliability range from .668 to .955 across grades and subjects, with the median estimate equal to .832. Item response theory estimates of subject scaled score reliability are not yet available, but they are likely to be higher than the cluster score reliability estimates. Raw agreement indices showing the reliability of Key Skill mastery classifications range from .515 to to .909 across grades and subjects; most indices are between .60 and .70.

References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Appendix B

ITBS/TAP Subtests

Grade	Reading	Language	Mathematics	Science	Social Studies
2	Pictures Sentences Stories	Spelling Capitalization Punctuation Usage & Expression	Concepts Problems Computation		
3 – 8	Reading Comprehension	Spelling Capitalization Punctuation Usage & Expression	Concepts Problems Computation	Science	Social Studies
9 – 10	Reading Comprehension		Mathematics	Science	Social Studies

MISSOURI MASTERY AND ACHIEVEMENT TESTS
INDIVIDUAL STUDENT REPORT, SPRING 1987

Please turn the page over for an explanation of these scores

Name: SAM COLLINS
Building: PLUM RIPE ELEMENTARY
District: GOOD SOIL RURAL DISTRICT #1
Dist. Code 997-789-3000
Test Date: SPRING 1987

Subject: READING/LANGUAGE ARTS
Level Form: 6/A Grade: 6

Estimated Comparable
National Percentile Ranking: 72 READING
57 LANGUAGE

	Student's Score	District Average	State Average
Subject:			
READING/LANGUAGE ARTS	317	290	300
Clusters:			
READING	339	292	300
LANGUAGE ARTS	312	283	300
WRITING	296	301	300

Key Skills:

Mastered

- B-1 Contextual Word Meaning
- B-2 New Word Meaning
- B-4 Synonyms/Antonyms
- C-1 Story Sequence
- C-2 Author's Purpose
- C-3 Fact/Opinion
- C-4 Cause-effect
- C-5 Character Comparison
- C-6 Main Idea
- C-7 Summarization
- C-8 Outcome Prediction
- C-9 Conclusions/Generalizations
- C-10 Story Elements
- C-11 Point of View
- D-2 Maps/Charts/Tables
- D-5 Appropriate Sources
- G-3 Effective Writing
- G-5 Spelling
- G-6 Capitalization

Not Mastered

- C-12 Figurative Language
- D-1 Learning Resources
- D-6 Directions
- G-4 Draft Revision
- G-7 Punctuation
- G-8 Grammatical Usage

NOTE – [illegible]

Appendix D

Chapter I Eligibility Standards

Grade	Percentile Rank
K-3	45
4-6	40
7	36
8	32
9	28
10	24

Note: Students scoring at or below the percentile rank are eligible to receive Chapter I services.

MISSOURI MASTERY AND ACHIEVEMENT TESTS

CHAPTER 1 ELIGIBILITY LIST

	ELIGIBLE IN 3 SUBJECTS	READING PERCENTILE	READING NCE	LANGUAGE PERCENTILE	LANGUAGE NCE	MATH PERCENTILE	MATH NCE
IEP	ADAMS JOHN	1	1	6	17	1	1
	ARTHUR CHERYL	10	23	4	13	9	22
	BRECKINRIDGE JOHN	5	15	21	33	6	17
	BURR ALICE	36	43	39	44	14	27
IEP	CALHOUN JOHN	2	7	1	1	14	27
	CLINTON GERRY	5	15	6	17	13	26
IEP	COLFAX SCHUYLER	1	1	10	23	1	1
	DALLAS SUE	8	20	10	23	1	1
	FAIRBANKS CHARLES	22	34	21	33	9	22
	FILMORE MILLIE	14	27	30	39	21	33
	GERRY ELBRIDGE	34	41	4	13	13	26
	HAMLIN HEATHER	18	31	26	37	26	37
	HENDRICKS THOMAS	14	27	30	39	21	33
	HOBART LINDA	27	37	18	31	14	27
	JEFFERSON THOMAS	10	23	18	31	14	27
	JOHNSON ANN	11	24	2	7	4	13
	JOHNSON RICHARD	22	34	10	23	8	20
IEP	KING WILMA	2	7	3	11	2	7
	MORTON LEVI	18	31	15	28	39	44
	ROSSEVELT CARRIE	5	15	10	23	9	22
	SHERMAN JAMES	11	24	6	17	3	11
	STEVENSON BETTY	22	34	6	16	6	17
IEP	TOMPKINS DANIEL	6	17	1	1	1	1
IEP	TYLER MARY	6	17	4	13	11	24
	VAN BUREN MARTIN	14	27	3	11	1	1